Aggregation Process for Software Engineering

Authors: **Dr. Óscar Dieste Tubío**
Dr. Ramón García Martínez
M. Ing. Enrique Fernández Pignataro

Contents

1 Introduction

The number of experiments conducted in the field of software engineering (SE) has been increasing significantly for some years. These experiments cover the widest range of subjects from testing technique performance, requirements elicitation to programming language performance, etc. While the experiments do turn up interesting knowledge, they are generally small (they seldom use over 20 experimental subjects; see, for example, the experimental studies identified in [Davis, A.; et al; 2006]). For this reason, if the information they turn up is to be of any use, the results have to be aggregated to be able to arrive at findings backed by as much empirical evidence as possible.

The aggregation of experiments involves combining the results of several experiments analysing the behaviour of a specific pair of treatments in a particular context to get *a single final result*. The new result will be *more general and reliable than the individual results* and will be underpinned by a greater level of empirical evidence [Cochrane; 2008].

For the results of an aggregation process to be really reliable, this process must steer clear of bias related to experiment search and selection. Therefore, results aggregation processes are generally part of a systematic review (SR). A SR is a procedure that applies scientific strategies to make the process of compiling, appraising and aggregating relevant experimental studies about a subject more reliable [Goodman, C.; 1996]. The aggregation strategy that SRs use to combine the results of individual studies is meta-analysis.

Meta-analysis is a collective name referring to a set of statistical methods that aim to find a numerical result that is a representative summary of the results of the individual studies, thereby amounting to an improvement on the individual estimates.

Note that meta-analysis can only be applied if certain requirements are met, such as a minimum number of properly compiled and homogeneous experiments

[Gurevitch, J. and L.V. Hedges; 1993]. This will assure that the finding reached is really solid and reliable.

Whereas experiment aggregation is by no means new to sciences like psychology, education or medicine, it was not proposed as an alterative for generating SE knowledge until the mid 1990s [Basili, V.; et al; 1996]. Several authors have addressed the issue since then. For example, worthy of note is the SR procedure developed by [Kitchenham, B. ; 2004]. This procedure was the result of adapting the SR processes developed in medicine, considering several experiment quality levels in keeping with SE's present context. As in medicine, this procedure recommends the use of the weighted mean difference (WMD) statistical method [Hedges, L.; Olkin, I.; 1985] to aggregate the results.

There are other authors who have worked on applying SRs, e.g. the systematic review recently developed by [Dyba, T .; et al; 2007]. This SR identifies 11 experimental studies linked to pair programmer performance and combines their results using a standard meta-analysis method (applying the WMD method suggested in [Kitchenham, B. ; 2004]).

Now, not all the work linked to aggregation carried out in the SE field has been successful. Take, for example, [Banker & Keremer; 1989], [Shull, *et al*.; 2003], [Hu, Q.; 1997], [Wohlin *et al*.; 2003], [Juristo *et al*.; 2004] or [Jørgensen, M; 2004] where the authors did manage to develop the experiment search and selection procedures, but the combination of the experiments through meta-analysis turned out to be impracticable. The key obstacles to the application of meta-analysis in the present SE context are related to the following problems:

- Shortage of experiments, replications and homogeneity among experiments [Davis, A.; et al; 2006], [Miller, J; 2000]. The consequence is that the results of standard aggregation techniques, which are generally applied to a sizeable number of studies, are found to be wanting in all-important precision.
- Failure to apply experiment reporting standards. For example, [Burton, *et. al*., 1990] do not publish variances and [Crandall Klein, *et. al*., 1989] do not even report the means of the experimental results. Under these circumstances it is impossible to apply meta-analysis.
- Wide-ranging internal quality. For example, although [Burton, *et. al*., 1988] and [Crandall Klein, *et. al*., 1989] deal with the same research topic, there is a big discrepancy as regards study conception and make-up. This means that the studies cannot be considered replications and cannot be used for a process of meta-analysis-based aggregation. If they were, the heterogeneity analysis would actually invalidate the results.
- Non-standardization of the response variables. For example, the studies by [Agarwal, *et. al*., 1990] and [Woody, *et. al*., 1996] use different response variables to analyse the same aspects. This means that these experiments cannot be considered replications and they will not be able to be part of the aggregation process.

Now, SE is not the only science that suffers from a shortage of experiments and where experiment development is very costly. Take ecology, for example, where the costs and time it takes to evaluate the growth of some trees are enormous [Worn, B.;

et al; 2007]. Neither is SE the only science where the quality of the developed studies (generally confined to laboratory tests) is questionable, driving down the quality or reliability of the findings. Sciences that are much more committed to experimentation, like medicine for example, face the same problem [Guerra Romero, L.; 1996] [Shekelle, P.; et al; 2003].

This volume describes a set of aggregation techniques used in several different branches of science. It also sets out a strategy for applying these techniques together to aggregate experimental studies conducted within the field of SE and thereby generate *pieces of knowledge based on the best available evidence*.

2 Statistical Methods of Aggregation

In the field of statistics, all measures that express a general characteristic of a population, such as the mean or variance of the values that a variable takes in all the individuals of the population, are called *parameters* [Epidat; 2008]. The real value of such population parameters is usually unknown, because, to calculate it, all the individuals in the population would have to be observed. This is out of the question in most situations. In most cases, it will only be possible to observe a group (of varying size) of individuals rather than the whole population, i.e. a sample.

The information gathered in the sample data can be used to approximate the knowledge of the population to the value of its parameters. This is an inductive or inferential method of knowledge acquisition known as *statistical inference*. Its development was pioneered by [Neyman, J. and Pearson, E.; 1933]. This field now encompasses a broad collection of methods for reaching findings about the population parameters from the information expressed as observed data of a sample. Generally, inference methods are divided into two major categories:

- Parameter estimation methods,
- Hypothesis testing methods.

Generally, hypothesis testing methods are applied to a study analysing the behaviour of two or more treatments to find out whether or not the differences in the results of the treatments are significant. In contrast, the idea behind running an aggregation process is to get an improvement *index*, indicating how much better one treatment is than the other. Therefore, aggregation methods should be classed as parameter estimation methods rather than hypothesis testing methods, even though their results are used to determine whether one treatment is better than another.

Like any statistical parameter, the improvement index estimated in aggregation processes cannot be analysed as a single result, as it is known to be an approximate estimate. Therefore, both the size of the effect and its respective confidence interval will have to be estimated [Gardner M; Altman D.; 1992].

The bounds of a confidence interval represent the range of values with a given probability of including the exact value of the parameter (the researcher sets this confidence level generally using values like 0.90, 0.95 or 0.99 generally).

As with most statistical methods, there are two clearly separate groups of aggregation methods [Hedges, L.; Olkin, I.; 1985]: *parametric* and *non-parametric* methods.

Parametric methods [García, R; 2004] are applied to evaluate a set of statistical variables (called parameters) and find out what the population is like, whereas non-parametric methods [García, R; 2004] make no hypotheses about the nature of the population.

Knowledge of what the behaviour of the population is like increases the precision of the estimated parameters. This is known as statistical power and is the key advantage that the parametric methods have over the non-parametric approaches.

A parametric method cannot be applied unless it is possible to determine at least whether the analysed populations have normal distributions or homogeneity of variance. But to be able to establish the population's distribution or homogeneity of variance, either the samples taken must have a minimum size (generally greater than 300 experimental subjects) [García, R; 2004]) or the researchers must be fully acquainted with the behaviour of the phenomenon and be able deduce these points. If researchers are unable to clearly determine what the distribution of the phenomenon is, they will have to use non-parametric aggregation methods, even though they are less statistically powerful than their parametric counterparts.

Chapters 3 and 4 describe two parametric and two non-parametric methods in detail.

3 Parametric Methods

As mentioned earlier, parametric methods are applied by evaluating a set of parameters to determine what the population is like. Generally, it is necessary to make sure that [Hedges, L.; Olkin, I.; 1985]:

- The samples were taken at random,
- The samples were taken independently,
- The populations are normally distributed, and
- The populations have homogeneity of variance.

Apart from the above aspects, a condition of parametric aggregation methods is that the response variables reported in the experimental studies to be aggregated should be checked for compatibility. For example, it is not valid to lump together in one aggregation process studies that analyse program size in lines of code with others that analyse size based on the number of bytes that the program occupies on the hard disk. Even so, it is valid for the studies to use different scales of the same variable, such as lines of code measured in units or in thousands.

In the following we present the two parametric aggregation methods that will be described in this chapter:

- ***Weighted Mean Difference (WMD)***, widely applied in medicine [Cochrane; 2008] and already applied in SE [Dyba, T.; et al; 2007],
- Parametric ***Response Ratio*** (parametric RR) recently developed within the field of ecology [Gurevitch, J. and L.V. Hedges, 1993] and not yet used in SE.

The following sections detail each of the above methods.

3.1 Weighted Mean Difference

The weighted mean difference technique (WMD) [Glass, G.; 1976] is the best known and most widespread technique for estimating effect size or how much better one treatment is than another for a treatment with continuous variables. This technique is conceptually simple: the *individual effect* estimator (the improvement rate of one treatment compared to another for each experiment) is estimated as the quotient of the mean differences and the standard deviation, and the *overall effect* (the general improvement rate of one treatment over another estimated by combining *n* experimental studies) is calculated as a weighted mean of the effect estimators of the individual studies.

In the following we describe how to estimate the individual and overall effect size.

3.1.1 Estimating the individual effect

Estimating the individual effect is to calculate, for a particular study, how much better the experimental treatment is than the control treatment. This is done by dividing the mean difference between the two treatments by the pooled variance [Glass, G; 1976]. The estimation function is as follows:

$g = \frac{Y^E - Y^C}{S_P}$ **(1)**	g is the effect size Y is the mean of the experimental (E) and control (C) groups S_p is the pooled standard deviation of both groups

Table 3.1: Estimating the effect size for one study

The pooled standard deviation mentioned in the above function is estimated by:

$S_P = \sqrt{\frac{(n^E - 1)(s^E)^2 + (n^C - 1)(s^C)^2}{n^E + n^C - 2}}$ **(2)**	S_p is the pooled standard deviation of both groups s is the standard deviation of the experimental (E) and control (C) groups n is the number of experimental subjects in the experimental (E) and control (C) groups

Table 3.2: Estimating the pooled standard deviation

N.B.
The results of applying function (1) can be divided into three possible types:

1. Positive, indicating that the experimental treatment is better than the control treatment,
2. Negative, indicating that the control treatment is better than the experimental treatment,
3. Zero, indicating that there is no difference between the experimental treatment and the control treatment.

We describe how to interpret these values later in section 3.1.3 (How to interpret results).

The effect estimation function (1) [Glass, G; 1976] was improved by [Hedges, L.; Olkin, I.; 1985]. They added a correction factor "J" to improve the accuracy of the results when there are not many experimental subjects (especially under 10). This is an important point when developing an aggregation process in SE, as experiments tend to have few experimental subjects at present (e.g. see [Burton, A.; et al; 1990.]). The new function (currently recommended by [Cochrane, 2008]) is as follows:

$d = J(N-2)\frac{Y^E - Y^C}{S_P}$ **(3)**	d is the effect size J (N – 2) = is the correction factor Y is the mean of experimental (E) and control (C) groups S_p is the pooled standard deviation of both groups N is the number of experimental subjects of both groups ($n_E + n_C$)

Table 3.3: Estimating effect size

The correction factor "J" can be taken from the following table.

m	J (m)	M	J (m)	m	J (m)	M	J (m)
2	0.5642	15	0.9490	27	0.9719	39	0.9806
3	0.7236	16	0.9523	28	0.9729	40	0.9811
4	0.7979	17	0.9551	29	0.9739	41	0.9816
5	0.8408	18	0.9577	30	0.9748	42	0.9820
6	0.8686	19	0.9599	31	0.9756	43	0.9824
7	0.8882	20	0.9619	32	0.9764	44	0.9828
8	0.9027	21	0.9638	33	0.9771	45	0.9832
9	0.9139	22	0.9655	34	0.9778	46	0.9836
10	0.9228	23	0.9670	35	0.9784	47	0.9839
11	0.9300	24	0.9684	36	0.9790	48	0.9843
12	0.9359	25	0.9699	37	0.9796	49	0.9846
13	0.9410	26	0.9708	38	0.9801	50	0.9849
14	0.9453						

Table 3.4: Hedges' correction values

where m = N – 2,

or estimated using the following function:

$J = 1 - \frac{3}{4m-1}$ **(4)**	J is the correction factor

Table 3.5: Estimating the correction factor

Having estimated the effect size, the confidence interval can be estimated by means of the following function [Hedges, L.; Olkin, I.; 1985]:

$d - Z_{\alpha/2}\sqrt{v} \le \lambda \le d + Z_{\alpha/2}\sqrt{v}$ **(5)**	d is the effect size Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally, 1.96 (α = 0.05) is used. v is the standard error

Table 3.6: Estimating the confidence interval

The function for estimating the standard error, mentioned in the above function, is [Hedges, L.; Olkin, I.; 1985]:

$v = \frac{\tilde{n} + d^2}{2(n^E + n^C)}$ **(6)**	v is the standard error ñ = (n^E+n^C) / (n^E*n^C) d is the effect size of individual studies n is the number of experimental subjects in the experimental (E) and control (C) groups

Table 3.7: Estimating the standard error

Example 3.1:

Suppose we have identified an experiment (see Table 3.8 for experiment data), and we want to find out whether the experimental treatment is better than the control treatment.

Y^E	Y^C	n^E	n^C	s^E	s^C
100	80	10	10	25	20

Table 3.8: Experiment results

Note 1:
Table 3.9 below details the meaning of the columns mentioned in Table 3.8.

Letter	Meaning
Y	Y is the mean of experimental (E) and control (C) groups
N	n is the number of experimental subjects in the experimental (E) and control (C) groups
S	S is standard deviation of the experimental (E) and control (C) groups

Table 3.9: Description of the variables in the experiment results table

As the number of experimental subjects is low, we will use Hedges' corrected function (3) to estimate the effect size. To do this, first we will use function (2) described above to estimate the pooled standard deviation:

$$S_P = \sqrt{\frac{(10-1)(25)^2 + (10-1)(20)^2}{10+10-2}} = 22.64$$

After estimating the pooled standard deviation, we have to estimate the correction factor. To do this, we will use function (4):

$$J = 1 - \frac{3}{4*18-1} = 0.96$$

From the data calculated above and the means, we will be able to calculate the effect size using function (3):

$$d = 0.96\frac{100-80}{22.91} = 0.84$$

Looking at the result, we can say that the experimental treatment is better than the control treatment because the estimated effect has a plus sign. To check this, let us estimate the confidence interval at a 95% confidence level.

To make function (5) easier to calcuate, first let us estimate ñ:

$$\tilde{n} = \frac{10+10}{10*10} = 0.2$$

After estimating "ñ" we will use function (6) to estimate the standard error (v):

$$v = \frac{0.2+0.84^2}{2(10+10)} = 0.218$$

Finally, we will apply function (5) to estimate the bounds of the interval at a 95% confidence level ($\propto = 0.05$) from the estimated data:

$$L_l = 0.84 - 1.96\sqrt{0.218} = -0.069$$
$$L_u = 0.84 + 1.96\sqrt{0.218} = 1.761$$

To give a clearer picture of the results, Figure 3.1 below illustrates the result graphically.

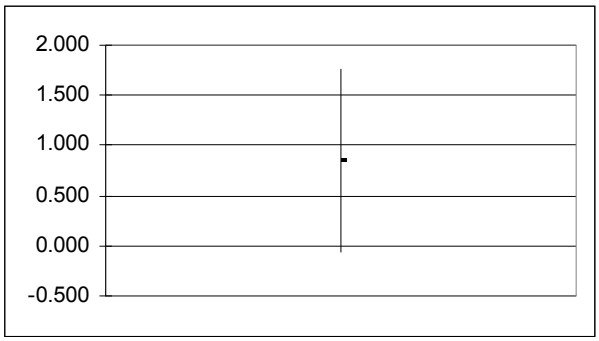

Figure 3.1: Graph of individual effect

As Figure 3.1 shows, the final estimated effect is positive and quite a long way from zero. This would suggest that the experimental treatment is better than the control treatment. But when we check this information against the confidence interval, we find that this interval contains zero. This indicates that the improvement of the experimental treatment over the control treatment is **not** significant at 95%.

3.1.2 Estimating of the overall effect

If all the studies included in the aggregation process were equally precise, it would suffice to estimate the effect size for each individual study and then average them out to get an overall effect size. In practice, though, not all the studies are equally precise. When combined, studies that supply more reliable information should be given more weight. A weighted mean is used to combine the results. A weighted mean differs from an ordinary mean, where all the elements are averaged based on their absolute value, in that each value is multiplied by a weight factor before being averaged [Borenstein, M.; et al; 2007].

As mentioned above, the overall effect is estimated as the weighted mean of the individual effects [Borenstein, M.; et al; 2007] [Hedges, L.; Olkin, I.; 1985], where each study is weighted as a function of the inverse variance. This way the more precise studies (generally studies including more experimental subjects) will be weighted higher because their results are considered to be more reliable or less error prone. The general estimation function is as follows:

$dw = w_1 * d_1 + ... + w_k * d_k$ **(7)**	dw is the overall effect size $w_1 \ldots w_k$ are the weight of individual studies $d_1 \ldots d_k$ are the effect size of individual studies

Table 3.10: Estimating the overall effect

Although each study is weighted based on the inverse variance, there are two views of how this variance should be estimated. These versions are known as the *fixed effect model* and the *random effects model*. The key differences between these two models are described below [Borenstein, M.; et al; 2007]:

Conceptually:

- The fixed effect model assumes that all the studies included in the aggregation process share one effect size value. The differences in the results of the different studies are due to experimental error (random error).
- In contrast, the random effects model assumes that each experimental study has its own effect, and there is no one effect size common to all the experiments. This is due to the fact that there are small variations in the results that cannot be controlled in the course of the experiments. For example, one noise factor that is hard to control in the SE context is the competence of developers participating in the experiments (which can be influenced by length of experience, higher education level or type of companies in which they have worked, etc.).

Influence of the studies on overall effect:

- For the fixed effect model, as all the studies are assumed to share the same effect size, the weights will be allocated entirely as a function of the quantity of information that each study offers. In practice, this will be reflected in the larger studies will be given a greater weight than the small studies.

- In contrast, the random effects model tries to estimate the mean of a set of effect sizes. Even though the the larger studies estimate more precise effects than the smaller studies, the weight of the larger studies will be less than for the fixed effect model, as we are looking for the mean of a set of real effect sizes rather than one real effect size.

Precision of the overall effect:

- For the fixed effect model, the only source of error is the random error that there is between the different studies. For this reason, if the final size of the sample is big enough, the error will tend to zero, irrespective of the fact that the aggregation process covers one study or many studies.
- In contrast, there are two error levels in the random effects model. The first is linked to the estimation of the effect sizes of each population of studies and the second is related to the estimation of the overall effect of combining all the individual effects. Therefore, the final precision of this model will depend on the quantity of subjects employed in each particular case and the number of experiments covered by the aggregation process.

In the following we describe how to estimate the overall effect for each model.

3.1.2.1 Fixed Effect Model

In this model [Borenstein, M.; et al; 2007] there is only one level of sampling, since all studies are sampled from a population with effect size μ. Therefore, we need to deal with only one source of sampling error – within studies.

Since our goal is to assign more weight to the studies that carry more information, we might set out to weight each study by its sample size. This way, a study with 1000 experimental subjects would get 10 times the weight of a study with 100 experimental subjects. This is basically the approach used, except that we assign weights based on the inverse of the variance rather than sample size. The inverse variance is roughly proportional to sample size, but has finer distinctions, and serves to minimize the variance of the combined effect.

In the following we detail how to estimate the overall effect. First, let us present the overall effect estimation function [Hedges, L.; Olkin, I.; 1985]:

$d^* = \frac{\sum d_i / \sigma^2{}_i(d)}{\sum 1/\sigma^2{}_i(d)}$ **(8)**	d^* is the overall size effect $\sum d_i / \sigma^2{}_i(d)$ is the sum of the individual effects $\sum 1/\sigma^2{}_i(d)$ is the sum of the inverse variance

Table 3.11: Estimating the overall effect

To estimate the overall effect, it is necessary to calculate the within-study variance ($\sigma^2(d)$). This is estimated as indicated below [Hedges, L.; Olkin, I.; 1985]:

$\sigma^2{}_i(d) = \frac{\tilde{n}_i + d^2{}_i}{2(n^E{}_i + n^C{}_i)}$ **(9)**	$\tilde{n}_i = (n^E{}_i + n^C{}_i) / (n^E{}_i * n^C{}_i)$ d_i is the effect size of individual studies n is the number of experimental subjects in the experimental (E) and control (C) groups

Table 3.12: Estimating the variance

After we have estimated the overall effect, we can use the following function to estimate the confidence interval of this effect [Hedges, L.; Olkin, I.; 1985]:

$d^* - Z_{\alpha/2}\sqrt{v} \le \lambda \le d^* + Z_{\alpha/2}\sqrt{v}$ **(10)**	d^* is the overall effect size Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally it is 1.96 ($\alpha = 0.05$). v is the standard error $(1/\sum 1/\sigma^2{}_i(d))$

Table 3.13: Estimating the confidence interval

Example 3.2:

Suppose that we have identified five experiments (see Table 3.14 for experiment data) and we want to find out whether the experimental treatment is better than the control treatment.

Id	Y^E	Y^C	n^E	n^C	s^E	s^C
1	100	80	10	10	25	20
2	100	105	8	8	15	14
3	100	85	20	20	12	12
4	95	100	4	4	15	15
5	110	75	20	20	18	16

Table 3.14: Experiment results

N.B.
The fields of Table 3.14 were described in Note 1 under Example 3.1.

Estimation of the overall effect size. Table 3.15 describes the results of applying the complementary functions making up function (8):

Id	Y^E	Y^C	n^E	n^C	s^E	s^C	d (3)	σ^2(d) (9)	$1/\sigma^2$(d)	d/σ^2(d)
E1	100	80	10	10	25	22.64	0.846	0.218	4.589	3.883
E2	100	105	8	8	15	14.51	-0.326	0.253	3.948	-1.286
E3	100	85	20	20	12	12.00	1.225	0.119	8.420	10.316
E4	95	100	4	4	15	15.00	-0.290	0.505	1.979	-0.574
E5	110	75	20	20	18	17.03	2.014	0.151	6.635	13.365
Total (8)									25.571	25.704

Table 3.15: Results of the complementary functions

Having estimated and summarized the effects and variances for each individual study, we can apply function (8) to estimate the overall effect:

$$d^* = \frac{25.704}{25.571} = 1.005 .$$

As we calculated the standard error (v) back in Table 3.14 (Column $1/\sigma^2(d)$), we can now apply function (10) to estimate the bounds of the confidence interval at a 95% confidence level ($\propto = 0.05$):

$$L_l = 1.005 - 1.96\sqrt{\frac{1}{25.571}} = 0.618$$

$$L_u = 1.005 + 1.96\sqrt{\frac{1}{25.571}} = 1.393 .$$

To give a clearer picture of the results, Figure 3.2 below illustrates the result graphically.

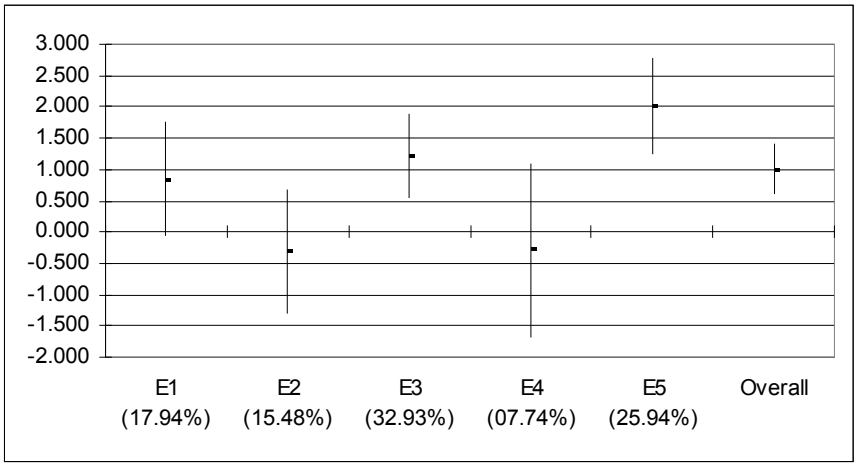

Figure 3.2: Graph of individual and overall effects

N.B.
The values between brackets underneath the identifier of each study

As Figure 3.2 shows, the final estimated effect is positive and quite a long way from zero. This would suggest that the experimental treatment is better than the control treatment. This can be checked by analysing the confidence interval, which does not contain zero. This indicates that the improvement of the experimental treatment over the control treatment is significant at 95%.

the individual studies. This is because the final results are substantiated by greater empirical evidence than the individual studies.

3.1.2.2 Random Effects Model

The fixed effect model [Borenstein, M.; et al; 2007] is based on the assumption that the true effect is the same in all studies. However, this assumption may be implausible in many systematic reviews. When we decide to incorporate a group of studies into a meta-analysis, we assume that the studies have enough in common for it to make sense to synthesize the information. However, there is generally no reason to assume that they are "identical" in the sense that the true effect size is exactly the same in all the studies. For example, the treatment might have a more pronounced impact in studies where the practitioners have more experience.

In the fixed effect analysis each study was weighted by its inverse variance. In the random effects analysis, too, each study will be weighted by its inverse variance. The difference is that the variance now includes the original (within-study) variance plus the between-studies variance, tau-squared.

In the following we describe how to estimate the overall effect. First, let us present the function for estimating the overall effect [Hedges, L.; Olkin, I.; 1985]:

$\Delta = \frac{\sum d_i / \gamma^2{}_i}{\sum 1/\gamma^2{}_i}$ **(11)**	Δ is the overall effect $\sum d_i / \gamma^2{}_i$ is the sum of the individual effects $\sum 1/\gamma^2{}_i$ is the sum of the inverse between- and within-study variances

Table 3.16: Estimating the overall effect

To estimate the overall effect, we should estimate the between-studies variance ($\sigma^2(\Delta)$), the within-study variance ($\sigma^2(d_i \mid \delta_i)$) and, from these, the overall variance (γ^2). Having estimated the variances, we will be able to estimate the overall effect.

To estimate the between-studies variance, we should apply the following function [Hedges, L.; Olkin, I.; 1985]:

$\sigma^2(\Delta) = s^2(d) - \frac{1}{k}\sum_{i=1}^{k}(c'_i + c''_i\, d^2{}_i)$ **(12)**	$\sigma^2(\Delta)$ is the between-studies variance $s^2(d)$ is the variance of the study effects k is the number of studies that are part of the aggregation process $C'_i = \frac{n_E + n_C}{n_E * n_C}$ $C''_i = \frac{a_i - 1}{a_i}$

	$a_i = \frac{(N_i - 2)\left[J(N_i - 2)\right]^2}{N_i - 4}$ d_i is the effect size of each study $N_i = n_E + n_C$ n is the number of experimental subjects in the experimental (E) and control (C) groups

Table 3.17: Estimating the between-studies variance

To estimate the variance between the effect sizes of the studies, the following function should be used [Hedges, L.; Olkin, I.; 1985]:

$s^2(d) = \sum_{i=1}^{k} \frac{(d_i - \overline{d})}{k-1}$ **(13)**	$s^2(d)$ is the variance of the study effects k is the number of studies that are part of the aggregation process d_i is the effect size of each study $\overline{d}$ is the average effect size of all the studies

Table 3.18: Estimating the variance of the effect sizes

The function for estimating the within-study variance is as follows [Hedges, L.; Olkin, I.; 1985]:

$\sigma^2(d_i \mid \delta_i) = (c'_i + c''_i\, d^2{}_i)$ **(14)**	$\sigma^2(d_i \mid \delta_i)$ is the within-study variance $C'_i = \frac{n_E + n_C}{n_E * n_C}$ $C''_i = \frac{a_i - 1}{a_i}$ $a_i = \frac{(N_i - 2)\left[J(N_i - 2)\right]^2}{N_i - 4}$ $N_i = n_E + n_C$ n is the number of experimental subjects in the experimental (E) and control (C) groups

Table 3.19: Estimating within-study variance

Now that we have estimated the between-studies and within-study confidence interval [Hedges, L.; Olkin, I.; 1985]:

$\gamma^2{}_i = \sigma^2(\Delta) + \sigma^2(d_i \mid \delta_i)$ **(15)**	γ_i^2 is the overall variance $\sigma^2(\Delta)$ is the between-studies variance (see function 12) $\sigma^2(d_i \mid \delta_i)$ is the within-study variance (see function 14)

Table 3.20: Estimating overall variance

In the following we describe how to estimate the confidence interval [Hedges, L.; Olkin, I.; 1985]:

$\Delta - Z_{\alpha/2}\sqrt{v} \le \Delta \le \Delta + Z_{\alpha/2}\sqrt{v}$	Δ is the overall effect size

(16)	Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally it is 1.96 ($\alpha = 0.05$). v is that standard error ($1/\sum 1/\gamma^2{}_i$)

Table 3.21: Estimating the confidence interval

Example 3.3:

Suppose that we have identified five experiments (see Table 3.22 for experiment data) and we want to find out whether the experimental treatment is better than the control treatment.

Id	Y^E	Y^C	n^E	n^C	s^E	s^C

Table 3.22: Results of experiments

First we are going to estimate the between-study variance. Table 3.23 describes the results of applying the complementary functions making up function (12):

Let us use the data from Table 3.23 to estimate the between-study variance (12):

$$\sigma^2(\Delta) = 1.014 - 0.235 = 0.780 .$$

Table 3.24 below describes the results of applying the complementary functions making up function (11).

								(14)	(15)	
E1	100	80	10	10	25	22.64	0.846	0.226	0.994	0.841
E2	100	105	8	8	15	14.51	-0.326	0.237	0.984	-0.321
E3	100	85	20	20	12	12.00	1.225	0.117	1.115	1.367
E4	95	100	4	4	15	15.00	-0.290	0.466	0.803	-0.233
E5	110	75	20	20	18	17.03	2.014	0.128	1.102	2.220
Total (11)									4.999	3.874

Table 3.24: Results of the complementary functions

Having estimated and summarized the effects and variances for each individual study, we can apply function (11) to estimate the overall effect:

$$\Delta = \frac{3.874}{4.999} = 0.775 .$$

As we calculated the estimated standard error (γ) back in Table 3.24 (column $1/\gamma^2$), we can apply function (16) to estimate the confidence interval bounds with a 95% confidence level ($\propto = 0.05$):

$$L_l = 0.775 - 1.96\sqrt{\frac{1}{4.999}} = -0.102$$

$$L_u = 0.775 + 1.96\sqrt{\frac{1}{4.999}} = 1.652 .$$

To give a clearer picture of the results, Figure 3.3 below illustrates the result graphically.

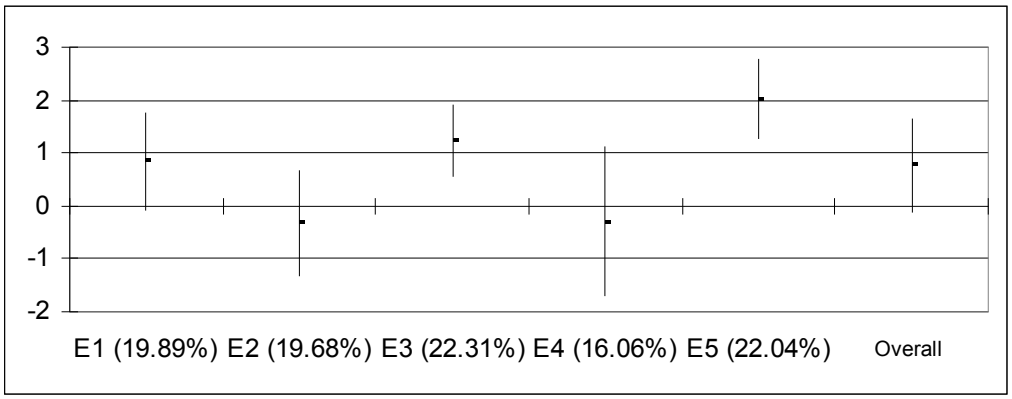

Figure 3.3: Graph of the individual and overall effects

N.B.
The values between brackets underneath the identifier of each study (E1...E5) indicate the weight that experiment carries in the final result. This is the result of dividing each of the values in column $1/\gamma^2$(d) of Table 3.24 by the sum total of this same column.

As Figure 3.2 shows, the final estimated effect is positive and quite a long way from zero. This would suggest that the experimental treatment is better

than the control treatment. But when we check this information against the confidence interval, we find that it contains zero. This indicates that the improvement of the experimental treatment over the control treatment is **not** significant at 95%.

3.1.2.3 Comparing Models

Table 3.25 is a brief comparison of the two models.

Aspect under evaluation	Fixed effect model	Random effects model
Similarity of the effect sizes in individual studies	Assumes that all the aggregated experimental studies share one effect size. Differences between individual effects are expected to be small.	Assumes that there is more than one effect size, because each study comes from a different population.
Statistical power	As it operates with one error level, the results tend to be more precise. This leads to narrower confidence intervals.	As it works with two error levels, the results tend to be less precise. This leads in broader confidence intervals.
Weight of individual experiments	As there is assumed to be just one effect size, the larger studies will be given more weight because they will be considered more precise.	As there is not assumed to be just one effect size, the weights of the studies will be more distributed, that is, the influence of the larger studies will be limited.
Influence of the number of experiments on precision	As there is assumed to be just one effect size and within study variance is the only estimated variance, this model is likely to be robust to the errors produced by a low number of experiments.	As there is not assumed to be just one effect size and both the within-study and the between-study variance have to be estimated, the between-study variance estimation error is likely to be high if there are not many experimental studies. For this reason, this model should not be used in such cases.

Table 3.25: Comparing the fixed effect and random effects model

3.1.2.4 What model should be applied in SE?

To answer this question, let us first highlight three important points:

1. Characteristics of experiments run in SE. Generally, aspects like productivity or analyst ability are not easy to estimate and control when evaluating the performance of a software engineering technique. As a rule, then, these aspects can be expected to generate some noise in the final findings.
2. Number of studies for aggregation. Currently, there are not many experimental studies that analyse the same treatments under the same experimental conditions. This means that there are not many experimental studies for aggregation in aggregation processes.
3. Size of experimental studies run in SE. Experimental SE studies are now usually run with few experimental subjects because they are costly in

monetary and development time terms. This means that the aggregation processes are generally run with studies of similar sizes.

Considering that the studies run in the current SE context are small, with few size differences and that the aggregation processes generally combine few experiments (rarely more than 10), *we recommend using the fixed effect model* rather than the *random effects model*, even though aspects like, for example, a developer's years of experience can be a difficult factor to control in some cases.

3.1.3 How to interpret the results

WMD-based meta-analysis provides a table for simply and clearly interpreting the results [Will Thalheimer and Samantha Cook]. This way, the end user can easily understand the generated pieces of knowledge. This table follows.

Effect size	*d*	Percentile standing	Per cent of non-overlap
	2.0	97.7	81.1%
	1.9	97.1	79.4%
	1.8	96.4	77.4%
	1.7	95.5	75.4%
	1.6	94.5	73.1%
	1.5	93.3	70.7%
	1.4	91.9	68.1%
	1.3	90	65.3%
	1.2	88	62.2%
	1.1	86	58.9%
	1.0	84	55.4%
	0.9	82	51.6%
LARGE	**0.8**	**79**	**47.4%**
	0.7	76	43.0%
	0.6	73	38.2%
MEDIUM	**0.5**	**69**	**33.0%**
	0.4	66	27.4%
	0.3	62	21.3%
SMALL	**0.2**	**58**	**14.7%**
	0.1	54	7.7%
	0.0	50	0%

Table 3.26: Interpreting the size effect

N.B.
The values presented in Table 3.26 are denoted with a plus sign, because the experimental treatment generally produces better results than the control treatment (or this is what the researcher is trying to demonstrate). However, the control treatment may happen to be better than the experimental treatment. In this case, the results will have a minus sign. Even so, they should be interpreted in the same absolute terms with respect to the ranges.

In the following we describe the meaning of each of the columns described in Table 3.26 [Cohen, 1988]:

- Effect sizes are hesitantly defined as "small, d = 0.2," "medium, d = 0.5," and "large, d = 0.8", stating that "there is a certain risk inherent in offering conventional operational definitions for those terms for use in power analysis in as diverse a field of inquiry as behavioural science".
- Effect sizes can also be thought of as the average percentile standing of the average experimental treatment to the average control treatment. An effect size of 0.0 indicates that the mean of the experimental group is at the 50th percentile of the control group. An effect size of 0.8 indicates that the mean of the experimental group is at the 79th percentile of the control group. An effect size of 1.7 indicates that the mean of the experimental group is at the 95.5 percentile of the control group.
- Effect sizes can also be interpreted in terms of the per cent of non-overlap between the experimental and the control groups' scores. An effect size of 0.0 indicates that the distribution of scores for the experimental group completely overlaps with the distribution of scores for the control group, i.e there is 0% non-overlap. An effect size of 0.8 indicates a non-overlap of 47.4% between the two distributions. An effect size of 1.7 indicates a non-overlap of 75.4% between the two distributions.

Example 3.4:

Looking at the effect sizes in examples 3.1 and 3.2, we can say that, as they are greater than 0.8 in both cases, the effects are large, and the experimental treatment is much better than the control treatment. Even though the effect is not greater than 0.8 for example 3.3, it is much closer to this value than to 0.2. For this reason, we are going to assume that the effect is also large.

Despite the fact that the final estimated effect is large, it is important to look at the confidence interval before generating the final conclusion. In this respect, the result of example 3.1 has a confidence interval ranging from –0.069 to 1.761. For this reason, even though the most likely effect is 0.84, we have to be careful about how we express this, because one of the possible results within a 95% confidence interval is 0. The same applies to the results of example 3.3 (the confidence interval is between –0.102 and 1.652). On the other hand, the confidence interval in example 3.2, from 0.618 to 1.393, is much narrower. In this case, we can be 95% confident that the effect size is large, that is, the experimental treatment is better than the control treatment.

3.1.4 Conclusions

- Pros
  - Hedges' corrected function minimizes the estimation error when studies are small (less than 10 subjects)
  - There are tables making it simple to interpret the results [Will Thalheimer and Samantha Cook]
  - It is known that this technique can be applied in aggregation processes of SE studies [Dyba, T.; et al; 2007]

- Cons
  - All the statistical parameters have to be published (which is not generally the case in the SE context)
  - The technique can only be applied if the experiment response variables are similar.

3.2 Response Ratio (Parametric)

The response ratio is the second parametric method that we are going to introduce. While this method is not as well-known as WMD, it is now the method ecologists recommend for aggregation processes within this field [Word, B.; et al; 2007] [Gurevitch, J. and Hedges, L.; 2001]. One of the reasons why it is the preferred method is that it has a low error rate in aggregation processes involving few studies [Lajeunesse, M & Forbes, M.; 2003]. This would appear to make it a good method for SE at present.

The response ratio is conceptually very simple. It consists of the calculating quotient of the means of an experimental treatment and a control treatment to estimate an improvement index. This quotient estimates how much better one treatment is than the other [Gurevitch, J. and Hedges, L.; 2001][Miguez, E. & Bollero, G; 2005]. For example, a ratio of 1.3 will indicate that the experimental treatment is 30% better than the control group or a ratio of 0.7 will indicate that the control treatment is 30% better than the experimental treatment. Clearly, the basis of this index is extremely simple.

Method application is a two-step process. The first step is to estimate the ratio of each experiment (what we will term individual effect estimation). After estimating the individual effect, we can then calculate a weighted average of the individual ratios to estimate the overall ratio or effect (which represents the general improvement rate of one treatment over another by combining *n* experimental studies) by.

The steps of this method are described in detail in the following.

3.2.1 Estimating the individual effect

To estimate the response ratio of an individual study, the mean of the experimental treatment should be divided by the mean of the control treatment [Hedges, L.; et al; 1999], as shown below.

$RR = \frac{Y^E}{Y^C}$ **(17)**	RR is the response ratio Y is the mean of the experimental (E) and control (C) groups

Table 3.27: Estimating the RR

N.B.
There are three possible results of applying (18):

1. Greater than "1" indicates that the experimental treatment is better than the control treatment,
2. Less than "1" indicates that the control treatment is better than the experimental treatment,
3. Equal to "1" indicates that there are no differences between the experimental and the control treatment.

Section 3.2.3 describes how to interpret these results.

Whereas directly calculating the quotient of the two means provides an improvement index for one individual study, the natural logarithm was added to assure that the combination of a set of studies is more precise [Hedges, L.; et al; 1999] [Miguez, E. & Bollero, G; 2005]. This linearizes the results (changes in the denominator have a bigger effect on the RR than changes in the numerator, whereas, thanks to the logarithms' properties, the Ln (RR) affects the numerator and denominator similarly). It also normalizes their distribution, making it a good method for estimating small sets of experiments. The new estimation function is:

$L = Ln(RR)$ **(18)**	L is the natural logarithm of the response ratio RR is the response ratio (17)

Table 3.28: Estimating Ln (RR)

Having estimated the ratio, we can use the following function to estimate its confidence level [Gurevitch, J. and Hedges, L.; 2001] [Miguez, E. & Bollero, G; 2005]:

$l - Z_{\alpha/2}\sqrt{v} \le \lambda \le l + Z_{\alpha/2}\sqrt{v}$ **(19)**	L is the natural logarithm of the response ratio Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally it is 1.96 ($\alpha = 0.05$). v is the standard error

Table 3.29: Estimating the confidence interval

The function for estimating the standard error, mentioned in the above function, is as follows [Hedges, L.; et al; 1999]:

$v = \frac{S^{2E}}{n^E Y^E} + \frac{S^{2C}}{n^C Y^C}$ **(20)**	V is the standard error S^2 is the variance of the experimental (E) and control (C) groups Y is the mean of the experimental (E) and control (C) groups n is the number of experimental subjects in the experimental (E) and control (C) groups

Table 3.30: Estimating the standard error

Having estimated the confidence interval, the antilogarithm should be applied to get the ratio index again.

N.B.
Using logarithms to convert the ratio used to estimate the confidence interval skews the interval bounds transformed by the antilogarithm.

Example 3.5:

Suppose that we have identified five experiments (see Table 3.31 for experiment data) and we want to find out whether the experimental treatment is better than the control treatment.

Y^E	Y^C	n^E	n^C	s^E	s^C
100	80	10	10	25	20

Table 3.31: Experiment results

N.B.
The fields of Table 3.31 were described in Note 1 under Example 3.1.

First let us estimate the RR based on function (17):

RR = 100 / 80 = 1.25

After calculating RR, we linearize the results applying function (18):

$L = Ln(1.25) = 0.223$

After estimating the Ln of the ratio, we can estimate the confidence interval. To do this, first we estimate the standard error (20):

$$v = \frac{25^2}{10*100} + \frac{20^2}{10*80} = 0.012 .$$

After estimating the standard error, we can apply function (19) to estimate the confidence interval bounds:

$$L_l = 0.223 - 1.96\sqrt{0.012} = 0.004$$
$$L_u = 0.223 + 1.96\sqrt{0.012} = 0.442 .$$

Finally, after estimating the confidence interval, we have to apply the

antilogarithm to get the original values and interpret the results:

A_{ll} = 1.004
L_u= 1.556

To give a clearer picture of the results, Figure 3.4 below illustrates the result graphically.

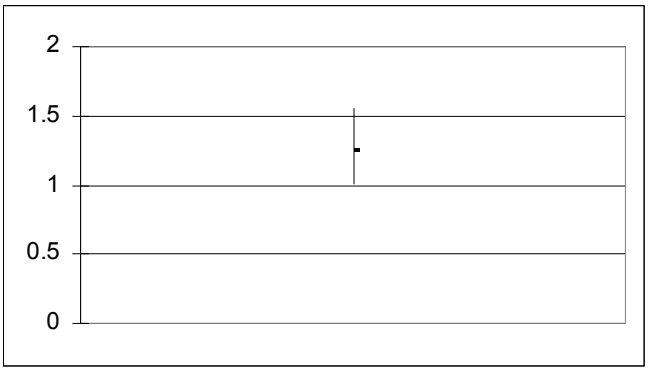

Figure 2.4: Graph of the individual effect

As Figure 3.4 shows, the final estimated effect is quite a long way from one. This would suggest that the experimental treatment is better than the control treatment. This can be confirmed by analysing the confidence interval, which does not contain 1. This indicates that we can be 95% confident the experimental treatment is better than the control treatment.

3.2.2 Estimating the overall effect

The overall ratio is estimated by the weighted average of the individual ratios [Gurevitch, J. and Hedges, L.; 2001], where each study is weighted depending on the inverse variance.

As the inverse variance tends to decrease as the number of subjects increases and this reduces the level of experimental error, the greater weights will be allocated to studies that include more experimental subjects because their results are considered to be more reliable or less error prone than the results of the small studies.

The overall ratio estimation function is described in the following.

$L^* = \frac{\sum_{i=1}^{k} W_i^* L_i}{\sum_{i=1}^{k} W_i^*}$ **(21)**	L^* is the overall effect L_i is the individual effect W_i is the weight factor = $1/v$ (where v is estimated as indicated in the function (20))

Table 3.32: Estimating the overall effect

After estimating the overall ratio, we will be able to estimate the confidence interval using the following function [Gurevitch, J. and Hedges, L.; 2001] [Miguez, E. & Bollero, G; 2005]:

$L^* - Z_{\alpha/2}\sqrt{v} \le \lambda \le L^* + Z_{\alpha/2}\sqrt{v}$ **(22)**	L^* is the overall effect Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally it is 1.96 ($\alpha = 0.05$). v is the standard error $(1/\sum 1/W_i))$

Table 3.33: Estimating the confidence interval

As in the case of the individual ratio estimation, the antilogarithm should be applied to the results after estimating the confidence interval to get the ratio index again.

Example 3.6:

Suppose that we have identified five experiments (see Table 3.34 for experiment data), and we want to find out whether the experimental treatment is better than the control treatment.

Id	Y^E	Y^C	n^E	n^C	s^E	s^C
1	100	80	10	10	25	20
2	100	105	8	8	15	14
3	100	85	20	20	12	12
4	95	100	4	4	15	15
5	110	75	20	20	18	16

Table 3.34: Experiment results

N.B.
The fields of Table 3.34 were described in Note 1 under Example 3.1.

First, let us estimate the overall ratio. To do this, Table 3.35 describes the results of applying the complementary functions making up function (21):

Id	Y^E	Y^C	n^E	n^C	s^E	s^C	RR (17)	L_i (18)	v (20)	W_i (21)	$L_i * W_i$
1	70	75	12	12	10	11	1.25	0.223	0.012	80	17.85
2	105	90	8	8	15	14	0.95	-0.048	0.005	198.62	-9.69
3	100	85	20	20	12	12	1.17	0.162	0.001	582.56	94.67
4	95	100	4	4	15	15	0.95	-0.051	0.011	84.33	-4.32
5	130	75	20	20	18	16	1.46	0.382	0.003	276.67	105.96
									Total (21)	1222.19	204.47

Table 3.35: Results of the complementary functions

Having estimated and summarized the effects and variances for each individual study, we can apply function (21) to estimate the overall effect:

$$L^* = \frac{204.47}{1222.19} = 0.167$$

As we calculated the estimated standard error (v) back in Table 3.35 (column

W_i), we can apply function (22) to estimate the confidence interval bounds with a 95% confidence level ($\propto = 0.05$):

$$L_i = 0.167 - 1.96\sqrt{\frac{1}{1222.19}} = 0.111$$

$$L_u = 0.167 + 1.96\sqrt{\frac{1}{1222.19}} = 0.223$$

To be able to correctly interpret the results of the mean effect and the confidence interval, let us now apply the antilogarithm to the results:

L* = 1.182
A_{ll} = 1.117
L_u = 1.250

To give a clearer picture of the results, Figure 3.5 below illustrates the results graphically.

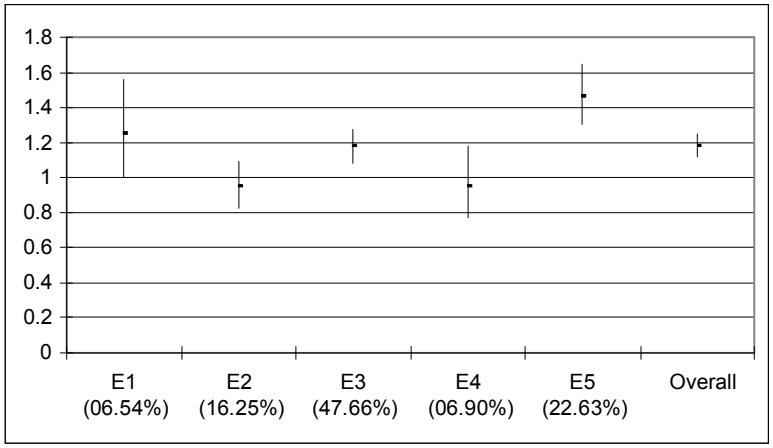

Figure 3.5: Graph of individual and overall results

N.B.
The values between brackets underneath the identifier of each study (E1...E5) indicate the weight that each experiment carries in the final result. This is the result of dividing each of the values in column W_i of Table 3.35 by the sum total of this same column.

As Figure 3.5 shows, the final estimated effect is positive and quite a long way from zero. This would suggest that the experimental treatment is better than the control treatment. This can be confirmed by analysing the confidence interval, which does not contain the one. This indicates that we can be 95% confident that the experimental treatment is better than the control treatment.

N.B.
The confidence interval obtained for the overall effect is quite a lot narrower than for the individual studies. This is partly due to the fact that the final result is founded on more empirical evidence than the individual studies.

3.2.3 How to interpret the results

In the case of the response ratio, there is no table for interpreting the results as there was for WMD. This time the results are analysed based on their absolute value, where a result equal to one means that the treatments are equivalent, a result greater than one means that the experimental treatment is better than the control treatment and a result of less than one that the control treatment is better than the experimental treatment [Gurevitch, J. and Hedges, L.; 2001] [Miguez, E. & Bollero, G; 2005]. For example, an effect of 1.25 indicates that the experimental treatment is 25% better than the control treatment.
Another important point to evaluate in the response ratio results is the confidence interval. To be able to claim with a particular confidence level that one of the treatments is better than the other, the confidence interval should not contain the value one. If it does, there will not be enough evidence, at the chosen significance level, to state that one of the treatments is better than another.

Example 3.7:

If we analyse the results in examples 3.5 and 3.6, we find in both cases that the resulting ratios are greater than one (1.25 and 1.18). This means that the experimental treatment is approximately 20% better than the control treatment. Additionally, as both confidence intervals are greater than one, we can be 95% confident that the experimental treatment is better than the control treatment in both cases.

3.2.4 Conclusions

- Pros
  - The error level of this method is low even if there are not many experimental studies for aggregation [Lajeunesse, M & Forbes, M.; 2003]
  - The confidence interval is narrow, which means that the result can be estimated very precisely

- No tables are required to understand the meaning of the final result

- Cons
  - All the statistical parameters have to be published
  - It is not known to have been applied to SE studies
  - The response variables of the experiments should be compatible

4 Non-Parametric Methods

As mentioned above, the non-parametric methods make no assumptions about the behaviour of the statistical parameters that affect the sample. Therefore, they are easier to apply, as they require only a minimum prior analysis of the phenomenon. However, they are less powerful, and more care has to be taken about expressing the results, as there is not as much knowledge of the phenomenon.

As it is better with statistics to make the mistake of saying that the results of two studies are equal when they really are not (type II error) than to make the mistake of saying that there are differences between the results when really there are not (type I error). The non-parametric methods apply a conservative policy that is reflected in the greater size of the estimated confidence intervals [Conover W; 1980]. This reduces the risk of the true result not being covered by the confidence interval.

The following are the non-parametric aggregation methods that we are going to analyse in detail in the following sessions:

- ***Vote Counting*** *(VC)* is a method that tries to estimate, based on the sign of the mean differences, the effect size that we would have measured using WMD if we had the necessary statistical parameters. We know of one aggregation in the field of SE so far that has used this method [Miller, J.; 2000]
- ***Response Ratio*****,** non-parametric version, (non-parametric RR), recently devised within the field of ecology [Worn, B.; 2007] and not yet used in SE.

4.1 Vote Counting

Vote counting is a method that can be applied even if there is very little information. All we really need to know is whether or not there is a mean difference between the treatments and how many experimental subjects each experimental study used. Although there are several versions of this technique, we will describe the version developed by [Hedges, L.; Olkin, I.; 1985] in this section. This version of the method is more than just a sum of votes, as its *objective is to estimate the effect size* (which we could have estimated if we had all the data required to apply WMD), based on the sign of the mean differences and the number of experimental subjects. An iterative inference process is applied to combine these parameters. This process aims to determine which is the most likely effect.

N.B.
This iterative inference process outputs reliable results when there are a sizeable number of studies and a similar number of experiments in favour of the experimental treatment and in favour of the control group. Otherwise there will be a tendency to overestimate the effect of the treatment that has more experiments in its favour [Lajeunesse, M & Forbes, M.; 2003].

Note importantly that when you apply this technique, the response variables included in the aggregation process should be related but do not need to be exactly the same. For example, it is valid for this method to combine experimental studies that measure the size of a programme in lines of code with another that measures size as a function of the hard disk space in MB occupied by the program. However, you have to be very careful with this, as the less alike the response variables in the aggregation process are, the greater the risk of reaching less reliable findings is.

As this method only measures the sign of the mean difference and the number of experimental subjects in each individual study, it is not possible to estimate the effect size for studies individually. For this reason, we only describe how to estimate the overall effect size (for the set of selected studies) below.

4.1.2 Estimating the overall effect

As mentioned above, the objective of this version of vote counting [Hedges, L.; Olkin, I.; 1985] is to estimate an effect size, like the WMD method does. In other words, it will not be confined to saying "the experimental treatment is better than the control treatment because it has more votes (studies in its favour)", but will estimate an improvement index that will determine how much better one treatment is than another.

To apply this technique, the first thing to do is set what is known as a "cut-off value" [Mohagheghi, P., & Conradi, R.; 2004]. This cut-off value indicates as of when the mean difference is considered to be valid for placing a vote. Generally, the difference should be significant at 0.05 (recommended by [Hedges, L.; Olkin, I.;

1985] to estimate a similar effect to the one estimated by WMD), but a lower cut-off value can be established, for example, to analyse whether the mean difference is greater than or less than zero.

N.B.
The less strict we are about defining the cut-off value, the greater the possibility of overestimation will be, as we are giving more votes to the main treatment than we really should. A similar thing could apply if, for example, a 99% confidence level were set for the differences between the treatments to be considered as a vote for the experimental treatment, whereas the confidence level for the final result of the aggregation test were set at 95%. In many cases, no significant differences would be found between the treatments when there really were some, and this would cause an underestimation of the effect size.

Having set the cut-off value, we should analyse each study to find out what category its "vote" belongs to. The possible categories to which the votes will be allocated are:

- Positive effect (the experimental treatment is better than the control treatment)
- Negative effect (the control treatment is greater than the experimental treatment)
- Zero effect (both treatments are equal)

For purposes of applying the method, there are only two values for the votes, 1 or 0. The positive effects belong to the first group and the negative and zero effects to the second.

After the votes have been placed, the next step is to carry out an iterative inference process based on maximum likelihood estimation (MLE). In this method, based on the votes and the number or experimental subjects in each study, we will determine what effect is mostly likely to occur within a range of preset effects. This range will extend from -0.5 to 0.5 as indicated in [Hedges, L.; Olkin, I.; 1985].

N.B.
Setting the bounds of the range of effects at -0.5 and 0.5 clearly shows that the authors were looking to develop an aggregation method to be applied in contexts where the number of votes for and against the treatment is similar, i.e. this method is not designed to detect big effects.

The function for establishing the likelihood of each effect is as follows [Hedges, L.; Olkin, I.; 1985]:

$$L(\delta \mid X_1,.....,X_i) = \sum_{i=1}^{k} \begin{Bmatrix} X_i \ln\left[1-\phi\left(-\sqrt{\tilde{n}}\delta\right)\right] + \\ (1-X_i)\ln\phi\left(-\sqrt{\tilde{n}}\delta\right) \end{Bmatrix}$$

$L(\delta|X_1,....X_n)$ is the probability of the effect
δ is the effect size to be tested

(23)	X_i is the value of each study's vote $\tilde{n} = (n^E+n^C) / (n^E*n^C)$ ϕ is the likelihood of the distribution being normal

Table 4.1: Estimating effect likelihoods

After establishing the most likely effect, we will be able to determine the confidence interval of this effect. This interval is generally broader than the one estimated by the WMD. The function for estimating the confidence interval is as follows [Hedges, L.; Olkin, I.; 1985]:

$\delta - Z_{\alpha/2} \sqrt{\overline{v(\delta)}} \le \lambda \le \delta + Z_{\alpha/2} \sqrt{\overline{v(\delta)}}$ **(24)**	δ is the most likely effect size Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally it is 1.96 ($\alpha = 0.05$). $v(\delta)$ is the standard error

Table 4.2: Estimating the confidence interval

Where the standard error ($v(\delta)$) is estimated as shown in the following function [Hedges, L.; Olkin, I.; 1985]:

$v(\delta) = \{\sum_{i=1}^{k} \frac{[D_i^1]^2}{p_i(1-p_i)}\}^{-1}$ **(25)**	k is number of experiments $p_i = 1 - \phi(-\sqrt{\tilde{n}_i}\delta)$ $D_i^1 = \sqrt{\frac{\tilde{n}_i}{2\pi}} e^{(-\frac{1}{2}\tilde{n}_i\delta^2)}$

Table 4.3: Estimating the standard error

N.B.
The results VC method can be divided into three possible types:

1. Positive, indicating that the experimental treatment is better than the control treatment,
2. Negative, indicating that the control treatment is better than the experimental treatment,
3. Zero, indicating that there are no differences between the experimental and the control groups.

We describe how to interpret these values later in section 4.1.2.

Example 4.1:

Suppose that we have identified five experiments (see Table 4.4 for experiment data), and we want to find out whether the experimental treatment is better than the control treatment.

Id	Y^E	Y^C	n^E	n^C	Significant
1	100	80	10	10	Unpublished
2	100	105	8	8	Unpublished
3	100	85	20	20	Unpublished
4	95	100	4	4	Yes
5	110	75	20	20	Yes

Table 4.4: Results of experiments

N.B.
Table 4.5 below describes the meaning of the columns mentioned in Table 4.4.

Initial	Meaning
Y	Y is the mean of the experimental (E) and control (C) groups
N	n is the number of experimental subjects in the experimental (E) and control (C) groups
Significant	Indicates whether a statistical test was used to test for the mean difference and this test indicated that the differences were significant

Table 4.5: Description of the variables in the experiment results table

Note also that Table 4.5 lists the same data as were used in Table 3.14, plus information about whether or not the mean differences are significant.

To apply the VC method, we first have to define the cut-off value. In this case, if we defined the cut-off value for comparing the means as a function of whether or not the differences are significant, we would only be able to compare two studies, because studies 1, 2 and 3 do not publish this information (this is commonplace in SE). Therefore, we will set a cut-off value based on the sign of the mean difference even though this puts the estimation at risk. Below we detail the result of the voting.

Id	Y^E	Y^C	n^E	n^C	Vote
1	100	80	10	10	1
2	100	105	8	8	0
3	100	85	20	20	1
4	95	100	4	4	0
5	110	75	20	20	1

Table 4.6: Voting

Now that we have established the votes, let us estimate the likelihood of each effect (23). To do this, we first detail how to estimate the likelihood for an individual effect (Table 4.7): an effect of 0.5 in this case. Then we present the likelihoods for the effects ranging from -0.5 to 0.5 (in Table 4.8).

Id	Y^E	Y^C	n^E	n^C	X	δ	$X_i \ln[1-\phi(-\sqrt{\bar{n}}\delta)]$ (23)	$(1-X_i)\ln\phi(-\sqrt{\bar{n}}\delta)$ (23)	$L(\delta\|X1,....Xn)$ (23)
1	70	75	12	12	1	0.5	-0.11691106	0	-0.11691106
2	105	90	8	8	0	0.5	0	-1.84102165	-1.84102165
3	100	85	20	20	1	0.5	-0.0586075	0	-0.0586075
4	95	100	4	4	0	0.5	0	-1.42815831	-1.42815831
5	130	75	20	20	1	0.5	-0.0586075	0	-0.0586075
Total (23)									-3.503

Table 4.7: Estimating the likelihood of effect 1

δ	L(δ|X1..X2)
0.5	-3.527
0.4	-3.228
0.3	-3.050
0.2	***-3.018***
0.1	-3.151
0	-3.465
-0.1	-3.977
-0.2	-4.698
-0.3	-5.638
-0.4	-6.806
-0.5	-8.205

Table 4.8: Result of estimating the likelihood of all effects

As Table 4.8 shows, the most likely effect is 0.2. As a result, we will say that the estimated size effect is 0.2. Now that we have established the effect size, we can use function (24) to infer the confidence interval. To do this, first we have to use function (25) to estimate the standard error.

Id	Y^E	Y^C	n^E	n^C	X	δ	P_i (25)	D_i^1 (25)	$v(\delta)$ (25)
1	70	75	12	12	1	0.2	0.672	0.807	3.614
2	105	90	8	8	0	0.2	0.655	0.736	2.818
3	100	85	20	20	1	0.2	0.736	1.032	8.200
4	95	100	4	4	0	0.2	0.611	0.542	1.339
5	130	75	20	20	1	0.2	0.736	1.032	8.200
Total (25)									24.172

Table 4.9: Estimating the complementary functions of the confidence interval

Now that we have estimated the standard error, we can apply function (24) to estimate the bounds of the confidence interval with a 95% confidence level ($\propto$ =0.05):

$$L_i = 0.2 - 1.96\sqrt{\frac{1}{24.172}} = -0.198$$

$$L_u = 0.2 + 1.96\sqrt{\frac{1}{24.172}} = 0.598$$

To give a clearer picture of the results, Figure 4.1 below illustrates the result graphically.

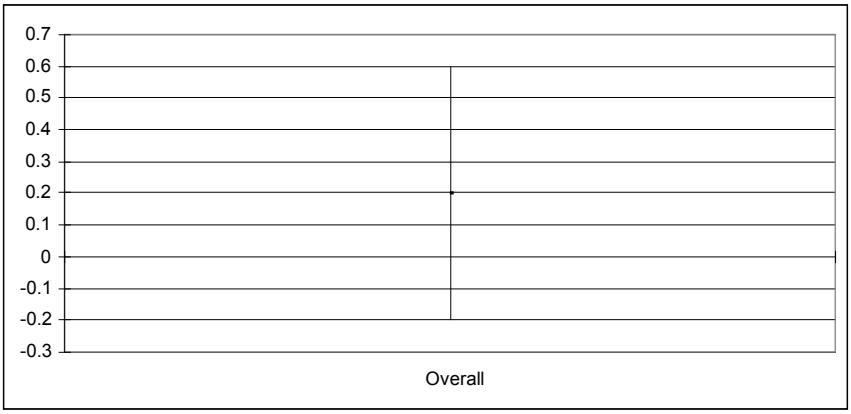

Figure 4.1: Graph of the individual effect

As Figure 4.1 shows, the final estimated effect is positive. This would suggest that the experimental treatment is better than the control treatment. However, when we check this information against the confidence interval, we find that it contains zero. This indicates that we can **not** be 95% confident that the experimental treatment is better than the control treatment.

4.1.3 How to interpret the results

As mentioned above, we can use this version of VC to estimate an effect size similar to WMD. For this purpose, it is valid to use the effect interpretation tables (Table 3.26), which we looked at under section 3.1.3. Table 4.10 below summarizes only the most significant values in that table.

d	effect size
0	None
0.2	Small
0.5	Medium
0.8	Large

Table 4.10: Interpreting effect size

Example 4.2:

If we analyse the results of example 8, we see that the result is 0.2. According to Table 4.10, this suggests a small effect. This means that the experimental treatment is slightly better than the control treatment. But as the confidence interval contains the value 0, the experimental treatment cannot be said to be better than the control treatment.

4.1.4 Conclusions

- Pros
  - It can be applied even if there are not many data
  - It can be used to evaluate more than one response variable together
  - It has been used within SE [Miller, J; 2000]
- Cons
  - The estimated effect error is high, especially when there are not many studies [Lajeunesse, M & Forbes, M.; 2003]
  - Homogeneity cannot be analysed (see section 3), as it is only possible to estimate the overall effect
  - It is harder to calculate as it is estimated through an iterative inference process
  - There are risks of overestimation or underestimation, especially if the significance levels applied to the "votes" do not match the significance level of the overall aggregation.

4.2 Response Ratio (Non-Parametric)

The second non-parametric method that we are going to present is the response ratio. This is a variation on the parametric response ratio presented in section 2.1.2. As mentioned earlier, the response ratio consists of calculating the quotient of the two means between the experimental and control treatments to estimate an improvement index. This quotient estimates the improvement rate between the two treatments [Gurevitch, J. and Hedges, L.; 2001][Miguez, E. & Bollero, G; 2005][Worn, B.; et al; 2007].

There is not much difference between the non-parametric and the parametric versions of the response ratio. The main difference is that this version of the method does not require any knowledge of how the population behaves (what class of distribution it has) and does not use the treatment variances to estimate the overall ratio. This makes it easier to apply in experimental environments where reporting is incomplete [Worn, B.; et al; 2007], as is now the case with SE (see, for example, [Crandall Klein, *et. al*., 1989]). Note that even though it is a non-parametric method, this technique's error level is low and, according to studies by [Lajeunesse, M & Forbes, M.; 2003], similar to the error of the parametric version.

As applies to the parametric version of RR, the response variables reported in the different experiments have to be compatible with each other for this technique to be applied. For example, it is not valid to lump together in one aggregation process studies that analyse program size as lines of code with others that analyse program size based on occupied hard disk space in bytes. It is valid, though, to put together studies that use different scales of the same variable, such as the number of lines of code measured in units or in thousands.

As mentioned in the section on the parametric RR, the application of the method is divided into two steps. First we have to estimate the ratio of each experiment (which we will call estimation of the individual effect). After this has been done, it is possible to estimate the overall ratio or effect (represents the overall rate of

improvement of one treatment over another after combining *n* experimental studies). To do this, we will calculate a weighted average of the individual ratios.

In the following we describe in detail the two steps that this method involves.

4.2.1 Estimating the individual effect

The response ratio is estimated by dividing the mean of the experimental treatment by the mean of the control treatment [Hedges, L.; et al; 1999] as shown below.

$RR = \frac{Y^E}{Y^C}$ **(26)**	RR is the response ratio Y is the means of the experimental (E) and control (C) groups

Table 4.11: Estimating the RR

N.B.
There are three possible results of applying function (26):

1. Greater than "1" indicates that the experimental treatment is better than the control treatment,
2. Less than "1" indicates that the control treatment is better than the experimental treatment,
3. Equal to "1" indicates that there is no difference between the experimental and the control treatment.

We describe how to interpret these results later, in section 4.2.3.

As we mentioned in our description of the parametric version of this method, the natural logarithm was added to improve the precision of function (26) [Hedges, L.; et al; 1999] [Miguez, E. & Bollero, G; 2005]. This way, the results can be linearized (Ln(RR) affects the numerator and denominator equally whereas RR is affected more by the changes in the denominator than in the numerator) and their distribution can be normalized, making this a good method for estimating small-sized experiments. The new estimation function is as follows.

$L_i = Ln(RR)$ **(27)**	L_i is the natural logarithm of Response Ratio RR is the Response Ratio (26)

Table 4.12: Estimating Ln(RR)

Having estimated the effect size, the confidence interval could be estimated by the following function [Gurevitch, J. and Hedges, L.; 2001] [Miguez, E. & Bollero, G; 2005]:

$L_i - Z_{\alpha/2}\sqrt{v} \le \lambda \le L_i + Z_{\alpha/2}\sqrt{v}$ **(28)**	L_i is the natural logarithm of the response ratio Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally it is 1.96 ($\alpha = 0.05$). V is the standard error

Table 4.13: Estimating the confidence interval

This version of the response ratio does not require knowledge of the variances to estimate the standard error as the original version does. Instead, it is estimated based on the number of subjects and the response ratio, as shown below [Worn, B.; et al; 2007]:

$v = \frac{n_C + n_E}{n_E n_C} + \frac{Ln(RR^2)}{2(n_C + n_E)}$ **(29)**	v is the standard error n is the number of experimental subjects in the experimental (E) and control (C) groups RR is the response ratio (see function 26)

Table 4.14: Estimating the standard error

Having estimated the confidence interval, we have to apply the anti-logarithm to the results to again get the ratio index. Note that a consequence of this is that the new confidence interval is not symmetrical.

Example 4.3:

Suppose that we have identified five experiments (see Table 4.15 for the experiment data) and we want to find out whether the experimental treatment is better than the control treatment.

Y^E	Y^C	n^E	n^C
100	80	10	10

Table 4.15: Experiment results

N.B.
The fields of Table 4.15 were described in Note 1 under Example 3.1.

First let us estimate the RR based on function (26):

RR = 100 / 80 = 1.25.

Now that we have the RR, let us apply function (27) to linearize the results:

$L_i = Ln(1.25) = 0.223$.

Now that we have estimated the Ln ratio, we will be able to estimate the confidence interval. To do this, let us estimate the standard error (29):

$$v = \frac{10+10}{10*10} + \frac{Ln(1.25^2)}{2(10+10)} = 0.211$$

After estimating the standard error, we can apply function (28) to estimate the bounds of the confidence interval:

$$L_l = 0.22 - 1.96\sqrt{0.211} = -0.677$$

$$L_u = 0.22 + 1.96\sqrt{0.211} = 1.123$$

Finally, to interpret the results we will apply the anti-logarithm to the estimated confidence interval:

$A_{ll} = 0.507$
$L_u = 3.076$

To give a clearer picture of the results, Figure 4.2 below illustrates the result graphically.

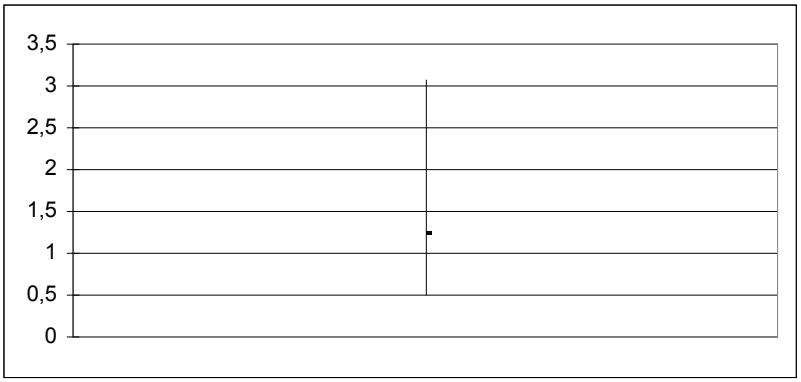

Figure 4.2: Graph of the individual effect

As Figure 4.2 shows, the final estimated effect is greater than one. This would suggest that the experimental treatment is better than the control treatment. But when we check this information against he confidence interval, we find that it contains one. This means that we can **not** be 95% confident that the experimental treatment is better than the control treatment.

4.2.2 Estimating the overall effect

We estimate the overall effect by calculating the weighted average of the individual effects [Curtis 1998], where each study is weighted according to its size (because the real variances are not known). This way, the studies that include more experimental results will have a greater weight as their results are considered to be more reliable or less error prone than the results from small studies. The estimation function is described below:

$L^* = \frac{\sum_{i=1}^{k} W_i^* L_i}{\sum_{i=1}^{k} W_i^*}$ **(30)**	L^* is the overall effect L_i is the individual effect W_i is the weight factor = $1/v$ (where v is estimated as indicated in function (29))

Table 4.16: Estimating the overall RR

Now that we have estimated the effect size, we can estimate the confidence interval using the following function [Gurevitch, J. and Hedges, L.; 2001] [Miguez, E. & Bollero, G; 2005]:

$L^* - Z_{\alpha/2}\sqrt{v} \le \lambda \le L^* + Z_{\alpha/2}\sqrt{v}$ **(31)**	L^* is the overall effect Z is the number of standard deviations that separate, at the specified significance level, the mean from the endpoint. Generally it is 1.96 ($\alpha = 0.05$). v is the standard error $(1/\sum 1/W_i))$

Table 4.17: Estimating the confidence interval

As we did to estimate the individual effect, we should apply, after estimating the confidence interval, the anti-logarithm to the results to again get the ratio index.

Example 4.4:
Suppose that we have identified five experiments (see Table 4.18 for the experiment data), and we want to find out whether the experimental treatment is better than the control treatment.

Id	Y^E	Y^C	n^E	n^C
1	100	80	10	10
2	100	105	8	8
3	100	85	20	20
4	95	100	4	4
5	110	75	20	20

Table 4.18: Experiment results

N.B.
The fields of Table 4.18 were described in Note 1 under Example 3.1.

First let us estimate the overall effect size. Table 4.19 describes the results of applying the functions making up function (30).

Id	Y^E	Y^C	n^E	n^C	RR (26)	L_i (27)	V (29)	W_i (30)	L_i*W_i
1	70	75	12	12	1.25	0.223	0.211	4.735	1.056
2	105	90	8	8	0.95	-0.048	0.246	4.049	-0.197
3	100	85	20	20	1.17	0.162	0.104	9.609	1.561
4	95	100	4	4	0.95	-0.051	0.493	2.025	-0.103
5	130	75	20	20	1.46	0.382	0.109	9.126	3.495
Total (30)								29.546	5.812

Table 4.19: Estimating the complementary functions for estimating the overall RR

Now that we have estimated and summarized the parameters for each individual study, we will be able to apply function (30) to estimate the overall effect:

$$L^* = \frac{5.812}{29.546} = 0.196 \ .$$

As we calculated the standard error (v) back in Table 4.19 (column W_i), we can now apply function (31) to estimate the bounds of the confidence interval at a significance level of 5%, that is,

$$L_l = 0.196 - 1.96\sqrt{\frac{1}{29.546}} = -0.163$$

$$L_u = 0.196 + 1.96\sqrt{\frac{1}{29.546}} = 0.557 \ .$$

To be able to correctly interpret the results of the mean effect and the confidence interval, let us now apply the anti-logarithm to the results:

$L^* = 1.217$
$A_{ll} = 0.848$
$L_u = 1.745$.

To give a clearer picture of the results, Figure 4.3 below illustrates the result graphically.

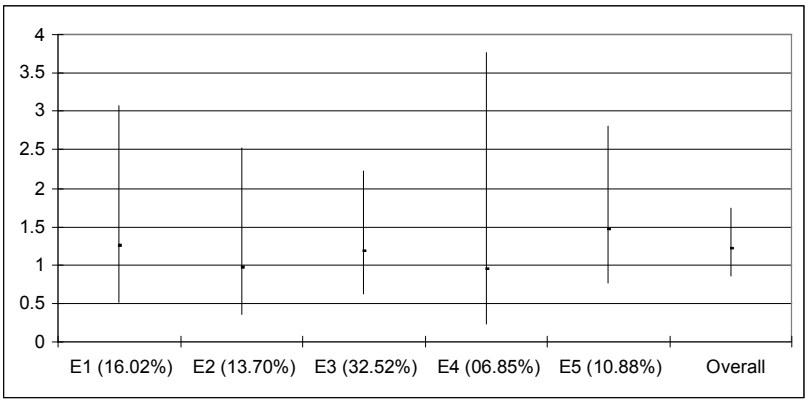

Figure 4.3: Graph of the individual and overall effects

N.B.
The values between brackets underneath the identifier of each study (E1...E5) indicate the weight that the experiment carries in the final finding. This is the result of dividing each of the values in column W_i in Table 4.19 by the sum total of this same column.

As Figure 4.3 shows, the final estimated effect is greater than one. This would suggest that the experimental treatment is better than the control treatment. But when we check this information against the confidence interval, we find

that the interval contains one. This means that we can **not** be 95% confident that experimental treatment is better than the control treatment.

N.B.
The confidence interval for the overall effect is quite a lot narrower than for the individual studies. This is partly due to the fact that the final result is founded by more empirical evidence than the individual studies.

4.2.3 How to interpret the results

As mentioned in section 3.2.3, there is no interpretative table for the response ratio as there is for the WMD. Here the results are analysed based on their absolute value, where a result that is equal to one means that the treatments are equivalent, a result that is greater than one means that the experimental treatment is better than the control treatment and a result of less than one means that the control treatment is better than the experimental treatment [Gurevitch, J. and Hedges, L.; 2001] [Miguez, E. & Bollero, G; 2005]. For example, an effect of 1.25 means that the experimental treatment is 25% better than the control treatment.

Another important point to evaluate in the response ratio results is the confidence interval. To be able to say with any level of confidence that one of the treatments is better than the other, the confidence interval should not contain the value one. If it does, there will not be evidence enough, at the chosen significance level, to say that either of the treatments is better than other.

Example 4.5:

Analysing the results of examples 4.4 and 4.5, we find in both cases that the ratios are greater than one. This means that the experimental treatment is approximately 20% better than the control treatment. But, as in both cases the confidence interval contains the value one, there is not enough evidence to say that the experimental treatment is better than the control treatment with a 95% confidence level.

4.2.4 Conclusions

- Pros
  - The error level is low even if there are not many experimental studies to be aggregated [Lajeunesse, M & Forbes, M.; 2003]
  - No tables are required to understand the final result

- No knowledge is required of how the population behaves (distribution or homogeneity of variance)
- Variances do not need to be known

- Cons
  - It has not yet been applied in SE
  - The experiment response variables should be similar
  - Being a non-parametric technique, the confidence intervals are greater than estimated by the parametric version.

5 Comparing Results

This chapter analyses the results of the aggregation techniques described in chapters 3 and 4. In this respect, we will first compare the results by the type of effect index they provide: effect size and response ratio. Then we will compare all the results.

5.1 Analysing the results of WMD and VC

To analyse the results, Figure 5.1 below shows the results of the estimated overall effects in examples 3.2, 3.3. and 4.1 together. This is possible as all the examples were calculated based on the same values.

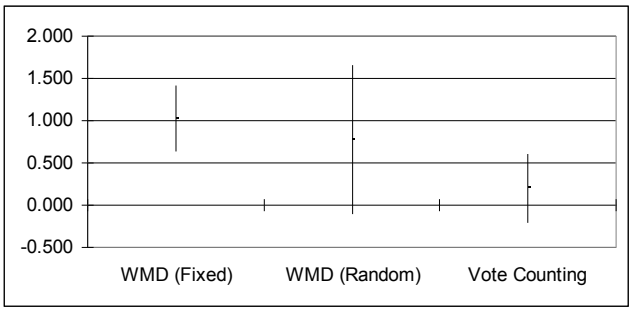

Figure 5.1: Graph of overall effects

From the results shown in Figure 5.1, we can say that:

- The confidence interval of the WMD fixed effect model is the narrowest. This indicates that this is the most precise technique.
- The confidence interval of the WMD random effects model is the widest. This is mainly due to the fact that the aggregation process was applied to few experimental studies. This increases the error level of the between-studies variance.
- VC estimates the smallest effect. This corroborates what we said about there having to be a lot of studies for the results of this technique to be precise, and

the number of studies in favour of the experimental and control treatments having to be similar.
- The fixed effect model is the only method that estimated significant differences at a 95% confidence level, as its confidence interval does not include zero.

Based on the above, we can say that, in this context, the best method is the fixed effect model, because it is highly precise. On the other hand, whereas the VC method did manage to detect that the effect was favourable to the experimental treatment, the effect it estimated was so far away from the random effects model estimate that the confidence intervals did not even overlap.

5.2 Analysing the results of parametric and non-parametric RR

To analyse the results, Figure 5.2 below illustrates the results of the overall effects estimated in examples 3.6 and 4.4 together. This is possible because all the examples were calculated based on the same values.

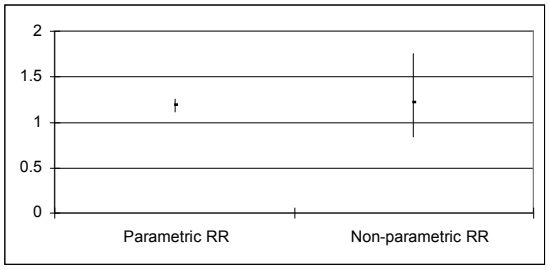

Figure 5.2: Graph of the overall effects

From the results shown in Figure 5.2 we can say that:

- The parametric RR resulted in an extremely narrow confidence interval. This indicates that the effect estimation is very precise even if there are not many studies.
- The parametric RR estimated significant differences at a 95%confidence level, as its confidence interval does not contain zero.
- The non-parametric RR estimates a very similar effect to the parametric RR, although its confidence interval is wider.

From the above, we can say that the parametric RR is a highly precise technique, whereas the non-parametric technique, being essentially more conservative, has need of more evidence to show up significant differences of effect.

5.3 Overall analysis

- Both the WMD fixed effect model and the parametric response ratio indicate that the experimental treatment is better than the control treatment at a 95% confidence level. On the other hand, the non-parametric techniques and the WMD random effects model turned up results indicating that the differences were not significant. This confirms the fact that the first two techniques are more powerful than the others.
- The non-parametric RR turned up quite similar results to the parametric version. This was not the case with VC, which, also being a non-parametric model, failed

to estimate results compatible with those estimated by the WMD fixed effect model.

From the above, we can say that the parametric techniques turned up results that were compatible with each other. This, however, did not apply to the non-parametric techniques. Within the second group, the RR technique turned up more reliable results than VC. This means that if there are not many studies, it is more reliable to use non-parametric RR.

6 Validating Results

6.1 Analysing Heterogeneity

When developing an aggregation process it is essential to make sure that the results of the experimental studies that were part of the aggregation process are compatible with each other. In other words, we have to check that the differences in the results of the studies are due to a random error in the experiment and not to an error caused by an uncontrolled external factor. This is known as *statistical homogeneity* and is evaluated through *heterogeneity analysis.*

There are several analytical and graphical methods for evaluating how heterogeneous a set of experimental studies are. These methods can assess the extent to which the results from different studies can be combined into a single measure.

Generally, all analytical tests designed to check for heterogeneity are based on the hypothesis of zero between-study variability. One of the best known tests for assessing statistical heterogeneity is the Q test proposed by [DerSimonian, R. and Laird, N.; 1986]. This test is generally recommended for reasons of validity and computational simplicity [Takkouche B.; et al; 1999]. Despite its pros, this analytical test is not very statistically powerful, especially when applied to a small number of experimental studies (as is usually the case in SE, where there are seldom more than 10 studies).

The idea behind this technique is that if there cannot be said to be heterogeneity, there is homogeneity. But the method is not very powerful if there are not many studies. Hence, it cannot be considered to provide evidence of homogeneity in this context, as it may fail to detect statistically significant differences in meta-analyses with moderate levels of heterogeneity [Epidat, 2008].

In view of the poor strength of evidence provided by existing analytical tests, graphical techniques can be used as an alternative way of checking for homogeneity

between the results of the experiments. Following advice by [Kitchenham, B; et al; 2004], we now describe how to use forest plots to represent the results of the different studies and how to check for homogeneity.

Interpreting Forest Plots

The forest plot [Molinero, L; 2006] is a graphical technique used to represent the results of meta-analyses. On the plot, the effect sizes (or ratios) of the individual studies and the overall effect size (or ratio) are represented differently. Below we detail how each result is represented:

- Individual effect size (or ratio):
  The confidence interval of each study is represented by means of a horizontal line, where a rectangle on the centre of the line is used to depict the individual effect size (or ratio).
- Overall effect size (or ratio):
  The confidence interval of the overall effect size is also represented by a horizontal line, but the effect size is represented by means of a diamond instead of the rectangle used for the individual effects.

Figure 6.1 below is an example of such a plot.

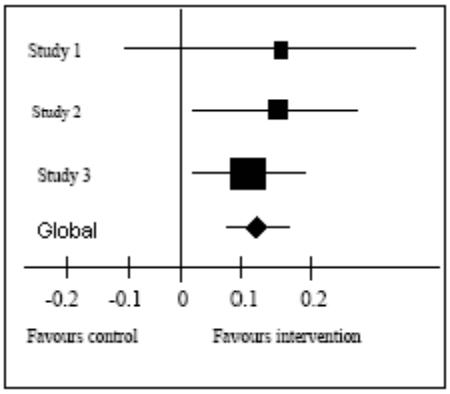

Figure 6.1: Example of Forest Plot

As Figure 6.1 shows, study 3 has a bigger impact on the results than studies 1 and 2 (the rectangle is much bigger). For this reason, the estimated overall effect is much more like this study than the other two evaluated studies.

As we are not familiar with any tool that jointly supports all the aggregation techniques represented here (WMD, parametric RR, vote counting and non-parametric RR), we proceeded to implement them using Microsoft Excel (version 2003) templates. Despite the wide variety of graphs, Microsoft Excel does not include a template for forest plots. Consequently, we have implemented this plot as follows.

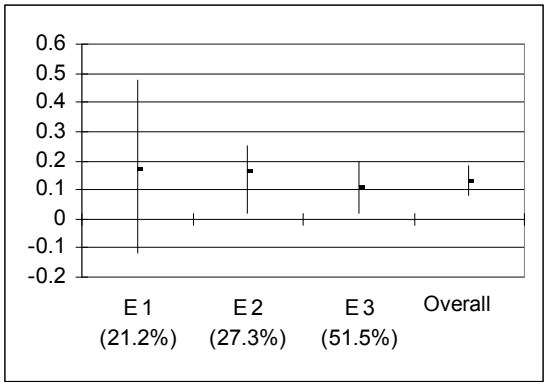

Figure 6.2: Representation of individual and overall effects

Unlike the standard diagram, the confidence intervals in the plot in Figure 6.2 are represented by vertical lines, and the impact of each study on the overall effect is detailed underneath the name of each experiment.

How to use the forest plot to check for homogeneity between studies

When the studies that are part of the aggregation process are homogeneous, the confidence intervals overlap with the confidence interval of the overall effect. If, on the other hand, the result of a study does not overlap with the confidence interval of the overall effect, this study will be said not to be homogeneous (or to be heterogeneous) with respect to the others.

Looking at the example in Figure 6.1, we can say that there is homogeneity between the studies, as the confidence intervals of the individual studies overlap with the overall confidence interval.

Note that if researchers find that a study is not homogeneous with the others, they should try to identify the reasons for this difference in the results. Apart from removing the heterogeneous study from the current aggregation process, this could entail undertaking new search processes of studies or performing new experiments.

6.2 When to evaluate heterogeneity

By definition any aggregation process intending to turn up reliable results should analyse heterogeneity to assure that its findings are really valid and universally understandable. For this to be possible, all the studies that are part of the aggregation process should analyse the same treatments using the same response variables and in a similar development environment.

Now, in the current stage of SE evolution, it is very hard to find a great many studies that use the same response variables to analyse the same treatments. Because of this, aggregation techniques like vote counting have emerged, where an approximate effect size can be estimated based on the sign of the mean differences without identifying the effect size of each of the studies in the aggregation process. On top of this, using this technique it is possible to combine studies that use different response variables. Consequently, we have to conclude that it is impossible

to estimate whether or not there is homogeneity between the results of each of the studies aggregated using this technique. The findings of this technique then are less reliable. This means that the pieces of knowledge gathered through this technique are of lower quality than knowledge gathered by techniques on which tests of homogeneity can be run (WMD, parametric and non-parametric RR).

7 Non-Statistical Aggregation

Whereas statistical aggregation is the best way of combining the results of the experiments identified in a SR [Cochrane; 2008], these techniques cannot always be applied in the current experimental context of SE (generally for reasons of non-standardization of the response variables and reporting quality problems), and a less formal and less reliable aggregation strategy has to be taken up if the process is to be performed at all (as was the case in [Davis, A.; et al;2006]). In this chapter, we will present an aggregation technique called direct vote counting (DVC) (not to be confused with the vote counting described in section 4.1). Basically, this technique consists of adding up the studies for and against each treatment and, based on those totals, determine which of the treatments is best. As there is more than one way of counting votes, we will follow some of the recommendations by [Mohagheghi, p.; et al; 2004] in this section to categorize the results of the different experimental studies.

Note importantly that when this technique is applied, the response variables included in the aggregation process do not have to be exactly the same. For example, it is valid to combine experimental studies that measure the program size in lines of code with another that measures size depending on the hard disk space in MB occupied by the program.

In the following we describe how this technique works.

7.1 Estimating the overall effect

The goal of this technique is to determine whether or not one experimental treatment is better than another. To do this, before the votes are allocated to each experiment, we have to determine the cut-off value as of which one of the treatments is considered have won the "vote". This cut-off value could be set, for example, at 51% of votes in favour. Note that the lower the cut-off value is, the greater the risk of reaching a mistaken finding is.

N.B.
To prevent claims of the cut-off value being accommodated to the results from invalidating the process, it must be set at the start of the aggregation.

Having set the cut-off value, the voting should take place by placing each study into one of the following five categories [Mohagheghi, p.; et al; 2004]:

1. Positive Vote with Evidence (for the experimental treatment): assigned when the mean difference is known to be significant at 0.05 (its symbol will be ++)
2. Positive Vote without Evidence (for the experimental treatment): assigned when we know no more than the mean of the experimental treatment is greater than the control treatment (its symbol will be +)
3. Zero Vote: assigned when the results of both treatments were the same (its symbol will be 0)
4. Negative Vote with Evidence (for the control treatment): assigned when the mean difference is known to be significant at 0.05 (its symbol will be --)
5. Negative Vote without Evidence (for the control treatment): assigned when we know no more than the mean of the control treatment is greater than the experimental treatment (its symbol will be -)

Generally, categories 1 and 2 are considered positive votes (for the experimental treatment), and categories 4 and 5 are considered negative votes (for the control treatment).

Having evaluated all the experiments and assigned the votes, the votes within each category will have to be counted and then expressed graphically to improve understanding. When we have done this, we have to find out if any of the treatments scored higher than the preset cut-off value, in which case it will be declared as the winning treatment. Otherwise, there will be said to be no difference between the treatments.

Example 7.1:

Suppose that we have identified five experiments (see Table 7.1 for experiment data), and we want to use direct vote counting to combine them and get an overall result.

Id	Y^E	Y^C	Significant
1	100	80	Unpublished
2	100	105	Unpublished
3	100	85	Unpublished
4	95	100	Yes
5	110	75	Yes

Table 7.1: Experiment results

N.B.
Table 7.2 describes the meaning of the columns used in Table 7.1.

Initial	Meaning
Y	Y is the mean of the experimental (E) and control (C) groups
Significant	Indicates whether a statistical test was used to test the mean differences and whether the test turned up significant differences

Table 7.2: Description of experiment results table

To apply this method, let us first define the cut-off value, which will be set at a minimum of 51% of positive votes and 20% of significant votes. This means that for one of the treatments to "win", it must have more than 50% of the positive votes (sum of + and ++) and at least 20% of the significant votes (++).

Having set the cut-off value, let us assign the votes to the different studies:

Id	Y^E	Y^C	Significant	Vote
1	100	80	Unpublished	+
2	100	105	Unpublished	-
3	100	85	Unpublished	+
4	95	100	Yes	--
5	110	75	Yes	++

Table 7.3: Results of voting

After the voting, we will count the votes, as shown in Table 7.4.

Vote category	Number of votes	%
++	1	20
+	2	40
0	0	0
-	1	20
--	1	20

Table 7.4: Vote counting

After analysing the data in Table 7.4, we can say that the positive votes ("++" and "+") total 60%, zero votes add up to 0% and the negative votes ("- -" and "-") sum 40%. To give a clearer picture of the results, Figure 7.1 below illustrates the result graphically:

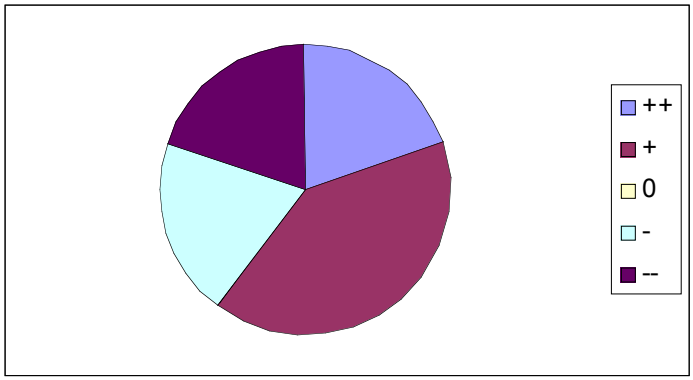

Figure 7.1: Percentage of votes

As Figure 7.1 shows, the majority vote was "+" and the positive votes together ("++" and "+") amount to more than 50% of the votes.

7.2 How to interpret the results

Direct vote counting can do no more than establish conjectures based on the percentage of votes counted; unlike the above statistical techniques, it does not provide effect or improvement indexes. Results are interpreted by saying whether the established cut-off values were exceeded after the votes had been counted.

Example 7.2:

Looking at the results in example 7.1, we find that they are above the preset cut-off value, as the positive votes ("++" and "+") are over 50%, and the significant positive votes are over 20% of the votes. Therefore, we will say that "the experimental treatment is presumed to be better than the control treatment".

7.3 Conclusions

- Pros
  - Not many data need to be known for it to be applied
  - It can be used to evaluate more than one response variable together
- Cons
  - No effect indexes can be established
  - The error level of the findings can be very high depending on the set cut-off factor
  - It has no statistical support
  - The homogeneity of the studies covered cannot be evaluated.

8 Case study of a real application

8.1 Introduction

This case study is based on the experimental studies identified in the systematic review process developed by [Davis, A.; et al; 2006]. The tasks of searching and evaluating studies were performed successfully in the above SR. But, due to reporting quality problems (most of the studies do not publish variances) and the shortage of studies that evaluate the same treatments jointly, the authors were unable to apply WMD in the aggregation process, as they had expected to. This meant that the only findings they were able to generate were based on direct vote counting. From this, the authors were able to infer that the structured interview technique appeared to gather more knowledge than the protocol analysis technique or that the laddering technique would gather similar amounts of knowledge to card sorting, for example.

As no procedure has yet been defined for systematically applying the aggregation techniques, they will be applied based on the characteristics of the study groups and their constraints.

8.1.1 Defining research questions

Before applying the aggregation techniques, we will define the experimental context in which they are to be applied. Table 5.1 below summarizes the key data from the viewpoint of the aggregation of the experimental studies selected for this aggregation.

Id	Evaluated techniques	Response variables	Reported statistics
1	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Number of rules ➢ Time taken ➢ Rule completeness	➢ Means ➢ Subjects
2	➢ Structured interview	➢ Number of clauses	➢ Means

Id	Evaluated techniques	Response variables	Reported statistics
	➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Time taken	➢ Variances ➢ Subjects
3	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Time taken	➢ Means ➢ Subjects
4	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Time taken	➢ Means ➢ Subjects
5	➢ Structured interview ➢ Twenty questions ➢ Card sorting	➢ Number of rules ➢ Implemented rules (%)	➢ Means ➢ Subjects
6	➢ Think-aloud ➢ Critical decision method	➢ Quantity of information gathered	➢ Mean difference ➢ Subjects
7	➢ Triads sorting ➢ Free sorting ➢ Direct sorting ➢ Ranking ➢ Picking from an attribute list	➢ Number of attributes	➢ Means ➢ Subjects
8	➢ Direct elicitation ➢ Rank ordering elicitation ➢ Ideal description	➢ Number of attributes	➢ Means ➢ Variances ➢ Subjects
9	➢ Direct elicitation ➢ Rank ordering elicitation ➢ Ideal description	➢ Number of attributes	➢ Means ➢ Variances ➢ Subjects
10	➢ Systematic interview ➢ Systemic interview	➢ Number of requirements	➢ Means ➢ Variances ➢ Subjects
11	➢ Open interview (*) ➢ Structured interview (*) ➢ Open interview (**)	➢ Number of rules ➢ Number of criteria	➢ Means ➢ Subjects
12	➢ Cognitive interview ➢ Standard interview	➢ Number of events	➢ Means ➢ Variances ➢ Subjects

Table 8.1: Details of studies for aggregation

(*) Knowledge engineers were novice
(**) Knowledge engineers were experts

N.B.
For more information about the studies described in Table 8.1, see Appendix A.

Before applying the aggregation methods, first let us present the possible groups of studies for aggregation. Then, based on these, we will specifically define the research questions. Looking at the data on the studies reported in Table 8.1, the groups were formed as follows:

1. Studies 1, 2, 3, 4 and 5 could be combined as they analyse the same techniques for eliciting decision rules.
2. Studies 7, 8 and 9 could be combined as they analyse several versions of the attribute elicitation technique.
3. Studies 9, 10 and 11 could be combined as they analyse several versions of the interview technique.
4. Study 6 would not appear to be able to be combined with other studies as it describes different elicitation techniques.

Now that we have established which studies can be aggregated with which, let us state the research questions to be answered by the aggregation process.

1) According to [Burton, A; et. al.;1988], there is more than one type of knowledge to be elicited during the construction of an expert system. There are two possible orders of knowledge:

- [a] first-order knowledge, which is the knowledge held by and handled by experts, and is what is usually meant by the term knowledge (models that the expert has about the world) and
- [b] second-order knowledge, which is knowledge that the knowledge engineer is looking for, and is knowledge about the expert knowledge, i.e. procedures, reasoning, heuristics, etc.

According to this definition the laddering and card sorting techniques are members of class [a], whereas structured interview and protocol analysis belong to group [b]. For this reason, we consider the best option to be a pairwise comparison of techniques depending on the type of knowledge that they can elicit.

The research questions are:

A) Is interviewing or protocol analysis better?

B) Is card sorting or laddering better?

2) Dynamic knowledge is a very important part of systems, but static knowledge is no less so. Direct elicitation and ranking [Bech-Larsen, T. et al., 1997] are the most important techniques used to elicit this type of knowledge. As these are the primary elicitation techniques for this type of knowledge, the two should be compared.

The research question is:

C) Is the direct elicitation technique better than ranking?

3) Although there are three studies that analyse the behaviour of different versions of the interview technique, these versions are incompatible with other. For this reason, we have not been able to specify a research question for this issue.

4) Whereas the think aloud and critical decision method might not appear to be compatible with the other techniques, both techniques tend, like interviews and protocol analysis, to gather second-order knowledge. In this respect, the basis of the think aloud technique has to do with experts expressing what they think verbally in a similar way to protocol analysis. A similar thing applies to the critical decision

method, where it is the expert that answers the questions asked by researchers as in interviews. As, at heart, these new techniques are the same as interviews and protocol analysis, the two types of techniques should be compared.

The research question is:

D) Are questioning-based techniques better than techniques of self-expression?

Now that we have defined the research questions, we have to define what response variables will answer these questions.

8.1.2 Defining response variables

Note that it is not an easy matter to determine whether one technique is better than another in the field of SE. In other branches of science a question asking whether a particular treatment is better than another can be dealt with directly (e.g. in medicine it could be enough to observe whether or not a patient recovers from a particular disease). But this case study observes and assesses many aspects about the analysed techniques, e.g. the number of rules inferred or the time taken in sessions or the complexity of applying the technique or the number of discovered attributes, etc.

To find out which the best response variables are, we analysed the selected studies. From this analysis, we determined that the response variables that best represented elicitation technique performance are:

- ***Gain***, calculated as the number of clauses, rules or attributes elicited in the sessions (depending on the analysed technique type).
- ***Effort***, calculated as the time it takes to carry out the requirements elicitation sessions.

8.2 Aggregating studies

Now that we have identified the experimental studies and set the variables for the questions to be answered, let us aggregate the studies to answer each defined research question.

8.2.1 First Aggregation

Question:

A) Is interviewing better than protocol analysis?

B) Is card sorting better than laddering?

Articles for aggregation:

Id	Evaluated techniques	Response variables	Reported statistics
1	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Number of rules ➢ Time taken ➢ Rule completeness	➢ Means ➢ Subjects
2	➢ Structured interview	➢ Number of clauses	➢ Means

Id	Evaluated techniques	Response variables	Reported statistics
	➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Time taken	➢ Variances ➢ Subjects
3	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Time taken	➢ Means ➢ Subjects
4	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Time taken	➢ Means ➢ Subjects

Table 8.2: Details of studies for aggregation

8.2.1.1 Defining the aggregation techniques for application

First, let us establish which aggregation techniques are applicable to the studies identified depending on the response variables and published statistics. As all the studies selected for this first aggregation analyse both pairs of techniques, we are going to present a single analysis to determine which aggregation techniques to apply to each study. If this were not the case, we would have to carry out a separate analysis for each different set of studies for aggregation.

To answer the questions we have studies 1, 2, 3 and 4. Looking at the statistical parameters published for these studies, we find that only study 2 publishes the standard deviations. This means that the parametric aggregation techniques cannot be applied. For this reason it will not be necessary to evaluate what the data distribution is like or check for homogeneity of variance. Following on with the analysis, as all these studies publish the means and the number of experimental subjects, the non-parametric aggregation and non-statistical aggregation techniques could be applied. Table 8.3 shows a summary of the techniques applicable to each study.

Id	WMD	Parametric RR	Non-parametric RR	Vote Counting	Direct Vote Counting
1			X	X	X
2	X	X	X	X	X
3			X	X	X
4			X	X	X

Table 8.3: Applicable aggregation techniques

Of all the analysed cases, study 2 can be aggregated by means of parametric techniques (WMD or RR). But this is the only study that can be aggregated by means of this technique type. As the aggregation of a single study does not input an aggregated value to the finding, parametric techniques will not be applied.

On the other hand, all the studies can be aggregated by means of non-parametric and non-statistical techniques. As the non-parametric RR is, for this group of techniques, the technique with the least error, it will be the technique used in all cases.

8.2.1.2 Applying aggregation techniques

In the following we will estimate the effect index for each of the defined research questions:

Applying non-parametric RR to "Structured Interview vs. Protocol Analysis" – "Gain"

Table 8.4 describes the individual data of each study for aggregation.

Id	Y^E	Y^C	n^E	n^C
1	94.4	75.8	16	16
2	274	145	16	16
3	270	269	4	4
4	317	184	4	4

Table 8.4: Experiment results

Table 8.5 shows the results of applying the non-parametric RR to the data in Table 8.4.

Id	RR	A_{II}	L_u	WEIGHT
1	1.245	0.611	2.537	41.568
2	1.889	0.896	3.984	37.829
3	1.003	0.250	4.015	10.952
4	1.722	0.393	7.546	9.649
Overall	1.469	0.928	2.325	

Table 8.5: Results of aggregation

Figure 8.1 is a graph of the results reported in Table 8.5.

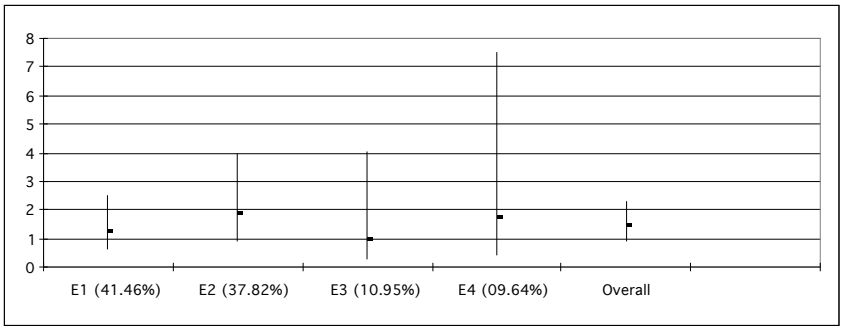

Figure 8.1: Results (global debe ser overall)

N.B.
The fact that the confidence intervals of the different studies overlap with the confidence interval of the overall effect corroborates that there is homogeneity between the results of the studies described in Figure 8.1.

Applying non-parametric RR to "Structured Interview vs. Protocol Analysis" – "Effort"

Table 8.6 describes the individual data of each study for aggregation.

Id	Y^E	Y^C	n^E	n^C
1	80.8	110.3	16	16
2	39.5	26.75	16	16
3	217	351	4	4
4	240	176	4	4

Table 8.6: Experiment results

Table 8.7 shows the results of applying the non-parametric RR to the data in Table 8.6.

Id	RR	A_{II}	L_u	PESO
1	0.732	0.376	1.425	43.17 %
2	1.476	0.714	3.051	36.27 %
3	0.618	0.168	2.268	11.31 %
4	1.363	0.323	5.747	9.23 %
Overall	0.981	0.633	1.519	

Table 8.7: Results of aggregation

Figure 8.3 is a graph of the results reported in Table 8.7.

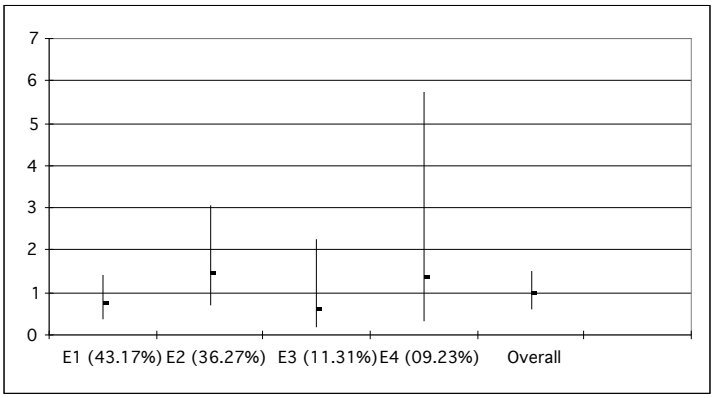

Figure 8.2: Results

N.B.
The fact that the confidence intervals of the different studies overlap with the confidence interval of the overall effect corroborates that there is homogeneity between the results of the studies described in Figure 8.1.

Applying non-parametric RR to "Card Sorting vs. Laddering" – "Gain"

Table 8.8 describes the individual data of each study for aggregation.

Id	Y^E	Y^C	n^E	n^C
1	63.4	101.4	16	16
2	420	521.4	16	16
3	188	123	4	4
4	278	216	4	4

Table 8.8: Experiment results

Table 8.9 shows the results of applying the non-parametric RR to the data in Table 8.8.

Id	RR	A_{II}	L_u	PESO
1	0,625	0,326	1,199	42,73%
2	0,806	0,411	1,580	39,87%
3	1,528	0,356	6,566	8,52%
4	1,287	0,308	5,373	8,87%
Overall	0,795	0,520	1,218	

Table 8.9: Results of aggregation

Figure 8.3 is a graph of the results reported in Table 8.9.

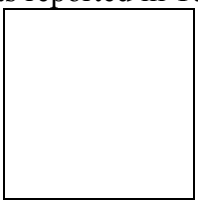

Figure 8.3: Results

N.B.
The fact that the confidence intervals of the different studies overlap with the confidence interval of the overall effect corroborates that there is homogeneity between the results of the studies described in Figure 8.3.

Applying non-parametric RR to "Card Sorting vs. Laddering" – "Effort"

Table 8.10 describes the individual data of each study for aggregation.

Id	Y^E	Y^C	n^E	n^C
1	67	79.8	16	16
2	29.75	40.75	16	16
3	145	98	4	4
4	177	145	4	4

Table 8.10: Experiment results

Table 8.11 shows the results of applying the non-parametric RR to the data in Table 8.10.

Id	RR	A_{II}	L_u	Peso
1	0,840	0,426	1,653	39,145

Id	RR	A_{ll}	L_u	Peso
2	0,730	0,375	1,420	40,630
3	0,676	0,181	2,521	10,375
4	0,819	0,212	3,163	9,850
Overall	0,773	0,506	1,182	

Table 8.11: Results of aggregation

Figure 8.4 is a graph of the results reported in Table 8.11.

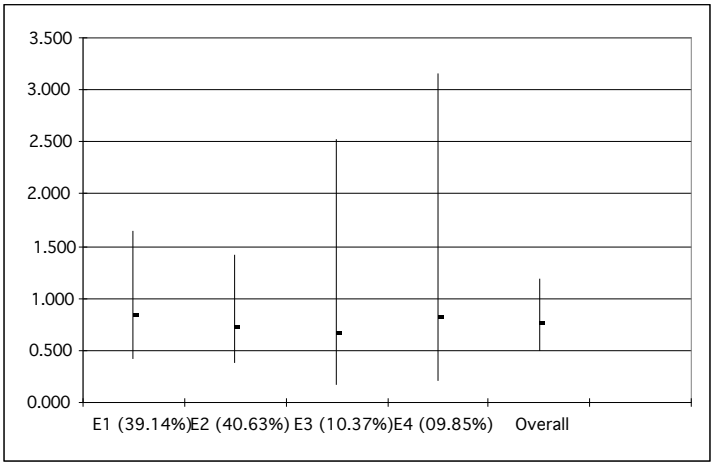

Figure 8.4: Results

N.B.
The fact that the confidence intervals of the different studies overlap with the confidence interval of the overall effect corroborates that there is homogeneity between the results of the studies described in Figure 8.4.

8.2.1.3 Analysing Results

Analysing the results of "Structured Interview vs. Protocol Analysis" — "Gain" and "Questioning vs. Self-expression" —"Gain"

1. The aggregation process combined four studies with altogether 80 experimental subjects.
2. The studies were aggregated using the non-parametric RR method. Being a conservative measure, the differences have to be sizeable for the confidence interval to indicate that the differences are significant.
3. Analysis of estimated results:
   - The final estimated ratio is **1.47,** which means that interviewing elicits almost **50% more rules** than protocol analysis.
   - The lower bound of the confidence interval is less than one. This means that there cannot be said to be a difference between the two techniques at a 95% confidence level. However, *as the non-parametric version of RR is fairly conservative in terms of confidence interval size, we would not be running much of a risk if we considered the interview technique to elicit more clauses than protocol analysis.*

4. Analysing the weights of the different studies, studies 1 and 2, being a lot bigger than the other two studies, can be said to carry most of the weight of the conclusion.
5. The results of all the studies can be said to be completely consistent. All the studies indicate that the experimental treatment is better than the control treatment.

Analysing the results of "Structured Interview vs. Protocol Analysis" — "Effort"

1. The aggregation process combined four studies with altogether 80 experimental subjects.
2. The studies were aggregated by means of the non-parametric RR method. Being a conservative measure, the differences have to be sizeable for the confidence interval to indicate that the differences are significant.
3. Analysis of estimated results:
   - The final estimated ratio is **0.98,** meaning that interviewing would take almost **2% less time** than protocol analysis.
   - As the confidence interval clearly includes the value one and the estimated effect is almost one, it can be said that there is no evidence to say that that either of the techniques is more efficient than the other with respect to administration of time.
4. Analysing the weights of the different studies, studies 1 and 2, being a lot bigger than the other two studies, can be said to carry most of the weight of the conclusion.
5. Consistency analysis: in this case, half of the studies support one technique and the other half the other by a wide margin.

Analysing the results of "Card Sorting vs. Laddering" — "Gain"

1. The aggregation process combined four studies with altogether 80 experimental subjects.
2. The studies were aggregated by means of the non-parametric RR method. Being a conservative measure, the differences have to be sizeable for the confidence interval to indicate that the differences are significant.
3. Analysis of estimated results:
   - The final estimated ratio is **0.795,** meaning that laddering can elicit a little over **20% more knowledge** than card sorting.
   - As the confidence interval clearly includes the value 1, we cannot be 95% confident that the laddering technique is better than card sorting.
4. Analysing the weights of the different studies, studies 1 and 2, being a lot bigger than the other two studies, can be said to carry most of the weight of the conclusion.
5. Consistency analysis: whereas half the studies support one technique and the other half the other technique, the studies with greater weight support the laddering technique, which is why the final effect favours this technique.

Analysing the results of "Card Sorting vs. Laddering" — "Effort"

1. The aggregation process combined four studies with altogether 80 experimental subjects.

2. The studies were aggregated by means of the non-parametric RR method. Being a conservative measure, the differences have to be sizeable for the confidence interval to indicate that the differences are significant.
3. Analysis of estimated results:
   - The final estimated ratio is **0.773,** meaning that *Laddering* takes just under 20% less time than card sorting.
   - As the confidence interval clearly includes the value 1, we cannot be 95% confident that the laddering technique is better than card sorting.
4. Analysing the weights of the different studies, studies 1 and 2, being a lot bigger than the other two studies, can be said to carry most of the weight of the conclusion.
5. Consistency analysis: the results of all the studies can be said to be consistent, as all the studies indicate that the experimental treatment takes less time than the control treatment.

General analysis

Based on the results from iterations 1 and 2, we can say that:

- The interview technique provides more knowledge than protocol analysis in a similar time.
- Although no significant differences were identified, the card sorting technique appears to take less time than laddering, but laddering also appears to elicit more knowledge than card sorting.

8.2.2 Second Aggregation

Question:

C) Is the direct elicitation technique better than ranking?

Studies for aggregation:

Id	Evaluated techniques	Response variables	Reported statistics
7	➢ Triads sorting ➢ Free sorting ➢ Direct sorting ➢ Ranking ➢ Picking from an attribute list	➢ Number of attributes	➢ Means ➢ Subjects
8	➢ Direct elicitation ➢ Rank ordering elicitation ➢ Ideal description	➢ Number of attributes	➢ Means ➢ Variances ➢ Subjects
9	➢ Direct elicitation ➢ Rank ordering elicitation ➢ Ideal description	➢ Number of attributes	➢ Means ➢ Variances ➢ Subjects

Table 8.12: Details of studies for aggregation

8.2.2.1 Defining the aggregation techniques to be applied

To answer the question we have studies 7, 8 and 9. If we analyse the statistical parameters published by these studies, we find that studies 8 and 9 publish all the

statistical parameters, and can be aggregated using parametric aggregation techniques. On the other hand, study 7 does not publish the standard deviations and can only be aggregated using non-parametric and non-statistical techniques. In summary, Table 8.13 shows the techniques applicable to each study.

Id	WMD	Parametric RR	Non-Parametric RR	Vote-Counting	Direct Vote-Counting
7			X	X	X
8	X	X	X	X	X
9	X	X	X	X	X

Table 8.13: Applicable aggregation techniques

Studies 8 and 9 can be aggregated by parametric techniques (WMD and RR), as well as less restrictive techniques. Hence all the studies can be aggregated by means of non-parametric and non-statistical techniques. Therefore, we will perform two aggregations. The first will include studies 8 and 9, and we will use parametric techniques and the second will include all three studies, and we will use the non-parametric RR. Non-parametric RR is the most reliable of the non-parametric and non-statistical techniques.

N.B.
Remember that, in this case, we will get findings of two different levels of reliability. The first will be formed by studies that can be aggregated by parametric techniques and the second by studies aggregated by the non-parametric RR.

8.2.2.2 Applying aggregation techniques

Applying WMD to "Direct Elicitation vs. Ranking" – "Gain"

Table 8.14 describes the individual data of each study for aggregation.

Id	Y^E	Y^C	n^E	n^C	S^E	S^C
8	4,49	4,32	43	39	1,4	2
9	4,95	4,85	43	39	1,6	1,83

Table 8.14: Experiment results

Table 8.15 below shows the results of applying WMD to the data in Table 8.14.

Id	ES	A_{II}	L_u	Weight
8	0.098	-0.335	0.532	49.98
9	0.058	-0.376	0.491	50.02
Overall	0.078	-0.228	0.385	

Table 8.15: Results of aggregation

Figure 8.3 is a graph of the results reported in Table 8.9.

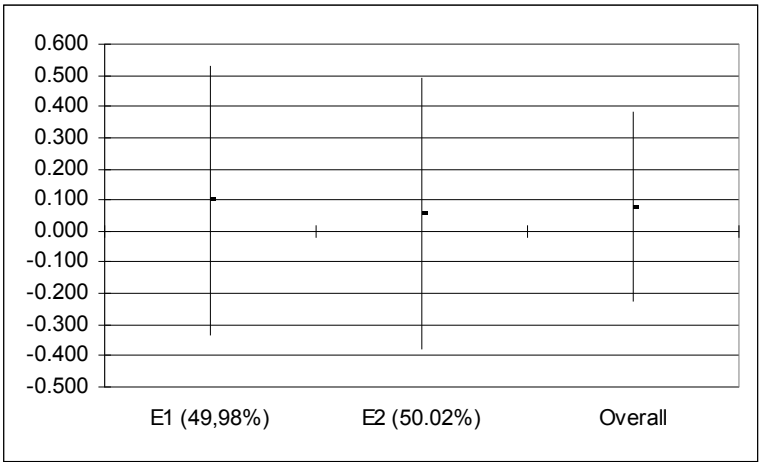

Figure 8.5: Results

N.B.
The fact that the confidence intervals of the different studies overlap with the confidence interval of the overall effect corroborates that there is homogeneity between the results of the studies described in Figure 8.5.

Applying parametric RR to "Direct Elicitation vs. Ranking" – "Gain"

Table 8.16 describes the individual data of each study for aggregation.

Id	Y^E	Y^C	n^E	n^C	S^E	S^C
8	4.49	4.32	43	39	1.4	2
9	4.95	4.85	43	39	1.6	1.83

Table 8.16: Experiment results

Table 8.17 below shows the results of applying parametric RR to the data in Table 8.16.

Id	RR	A_{ll}	L_u	WEIGHT
8	1.039	0.875	1.235	43.94 %
9	1.021	0.876	1.189	56.06 %
Overall	1.028	0.917	1.153	

Table 8.17: Results of aggregation

Figure 8.6 is a graph of the results reported in Table 8.17.

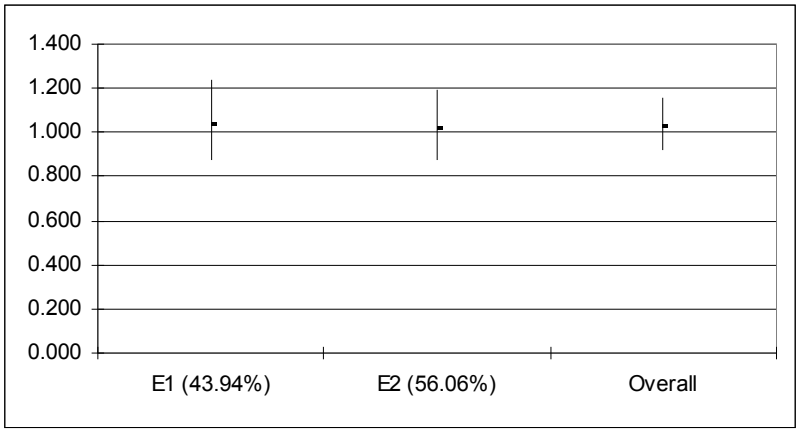

Figure 8.6: Results

N.B.
The fact that the confidence intervals of the different studies overlap with the confidence interval of the overall effect corroborates that there is homogeneity between the results of the studies described in Figure 8.6.

Applying non-parametric RR to "Direct Elicitation vs. Ranking" – "Gain"

Table 8.18 describes the individual data of each study for aggregation.

Id	Y^E	Y^C	n^E	n^C
7	8.60	9.53	30	30
8	4.49	4.32	43	39
9	4.95	4.85	43	39

Table 8.18: Experiment results

Table 8.19 below shows the results of applying parametric RR to the data in Table 8.18.

Id	RR	A_{ll}	L_u	WEIGHT
7	0.902	0.548	1.487	27.49%
8	1.039	0.672	1.607	36.17%
9	1.021	0.661	1.576	36.33%
Overall	0.993	0.764	1.290	

Table 8.19: Results of aggregation

Figure 8.7 is a graph of the results reported in Table 8.19.

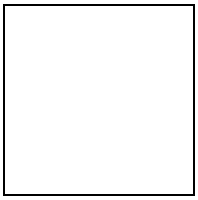

Figure 8.7: Results

N.B.
The fact that the confidence intervals of the different studies overlap with the confidence interval of the overall effect corroborates that there is homogeneity between the results of the studies described in Figure 8.7.

8.2.2.3 Analysing the Results

Analysing the results of "Direct Elicitation vs. Ranking" – "Gain"

1. The aggregation process included the following groups of studies:
   a. 2 studies with altogether 164 experimental subjects.
   b. 3 studies with altogether 224 experimental subjects.
2. The applied aggregation techniques were:
   a. WMD and parametric RR, which are the more precise aggregation techniques.
   b. Non-Parametric RR, which, being inherently conservative, requires quite sizeable differences for the confidence interval to indicate that the differences are significant.
3. Analysis of estimated effects:

Technique applied	Findings
WMD	➢ The estimated effect size is **0.078,** meaning that the improvement effect is almost zero. ➢ The confidence interval bounds clearly contain the value 0, confirming that there is no improvement effect in favour of either treatment.
RR Parametric	➢ The final estimated ratio is **1.028,** meaning that the improvement effect is almost zero (approximately 2%). ➢ The confidence interval bounds clearly contain the value 1, confirming that there is no improvement effect in favour of either treatment.
Non-Para-metric RR	➢ The final estimated ratio is **0.993,** which means that the improvement effect is almost zero (approximately 1%). ➢ The confidence interval bounds clearly contain the value 1, confirming that there is no improvement effect in favour of either treatment.

4. The weights of the different studies are well balanced.
5. The results of the different evidence levels can be said to be absolutely consistent. They all turn up a zero improvement level between the two treatments.

8.2.3 Third Aggregation

Question:

D) Are the questioning-based techniques better than techniques of self-expression?

Articles that can be aggregated:

Id	Evaluated techniques	Response variables	Reported statistics
1	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Number of rules ➢ Time taken ➢ Rule completeness	➢ Means ➢ Subjects
2	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Time taken	➢ Means ➢ Variances ➢ Subjects
3	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Time taken	➢ Means ➢ Subjects
4	➢ Structured interview ➢ Protocol analysis ➢ Card sorting ➢ Laddering	➢ Number of clauses ➢ Time taken	➢ Means ➢ Subjects
6	➢ Think-aloud ➢ Critical decision method	➢ Quantity of information gathered	➢ Mean difference ➢ Subjects

Table 8.24: Details of studies to be aggregated

8.2.3.1 Defining the aggregation techniques to be applied

Even though studies 1, 2, 3 and 4 were successfully aggregated under point 8.2.1, the addition of study 6 generates the following problems:

1. Response variable compatibility: Although study 6 indicates which of the techniques outputs more knowledge, it does not state whether this was measured as number of rules or clauses. This means that neither parametric aggregation techniques nor the non-parametric RR are applicable.
2. Publication bias: As study 6 only indicates that there is a mean difference but does not specify the specific mean values, none of the techniques requiring these values will be applicable.
3. Treatment compatibility: This aggregation focuses on combining the results achieved using essentially similar rather than the same treatments. This means that statistical techniques are not applicable. The only option for combining these results is to use direct vote counting, as no other techniques provide for such combinations.

The aggregation techniques that are applicable in this context are specified below.

Id	WMD	RR-Parametric	Non-Parametric RR	Vote-Counting	Direct Vote-Counting
1					X
2					X
3					X

Id	WMD	RR-Parametric	Non-Parametric RR	Vote-Counting	Direct Vote-Counting
4					X
6					X

Table 8.25: Applicable aggregation techniques

8.2.3.2 Applying aggregation techniques

Applying Vote Counting to "Questioning vs. Self-Expression" – "Gain"

Before voting takes place, we set the cut-off value at a minimum of 51% of votes in favour with 30% significant votes.

Table 8.26 below describes the individual data of each study.

Id	Significant	Vote
1	Unpublished	+
2	Published	++
3	Unpublished	+
4	Published	++
6	Unpublished	+

Table 8.26 Experiment results

Table 8.27 below shows the results of applying direct vote counting.

Vote category	Number of votes	%
++	2	40
+	3	60
O	0	0
-	0	0
--	0	0

Table 8.27: Results of aggregation

Figure 8.8 is a graph of the results reported in Table 8.27.

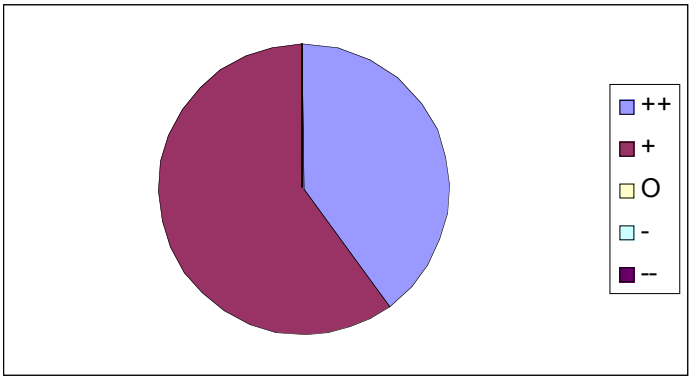

Figure 8.8: Results

8.2.3.3 3 Analysing the Results

Analysing the results of "Questioning vs. Self-Expression" – "Gain"

1. The aggregation process combined:
   a. 5 studies with altogether 50 experimental subjects.
2. The applied aggregation techniques were:
   direct vote counting, which, being a non-statistical technique, does not estimate effect sizes or confidence intervals. This knowledge will not be very reliable.
3. Analysis of estimated results:
   The established threshold values were surpassed: 30% of studies with significant differences and 50% of studies in favour. Therefore, questioning-based techniques are presumed to output more information than techniques in which the expert is expected to use self-expression.

8.3 Analysing results

In the following we summarize the most significant knowledge:

- Interviews elicit more rules than protocol analysis in the same time.
- Laddering elicit more rules than card sorting even though it takes longer per session.
- There do not appear to be performance differences between the attribute elicitation techniques of sorting and ranking.
- Questioning-based elicitation techniques elicit more knowledge than techniques based on expert self-expression.
- It was not possible to determine which version of the interview technique is better, as there is a problem of response variable incompatibility and very few replications of studies analysing the same versions of the technique.

8.4 Overall conclusions

As mentioned throughout this chapter, it is feasible to aggregate experimental studies in SE today, even though there are few studies and very often study reporting is not of good quality.

Nevertheless, care should always be taken about how the findings from both versions of vote counting are expressed.

9 Guidelines for applying aggregation techniques together

9.1 INTRODUCTION TO THE AGGREGATION PROCESS

Chapter 8 illustrated how the aggregation techniques described in chapters 3 (Statistical Aggregation Techniques) and 5 (Non-Statistical Aggregation Techniques) can be used depending on the features of the identified studies. No particular method or process was used to do this. Each technique has pros and cons, which are basically linked to the ease of application and level of precision of the response they output. This chapter presents a process of aggregation for applying techniques systematically. This should make it easier to output of pieces of knowledge based on the *best available evidence*.

Note, importantly, that using several aggregation techniques together to solve the "shortage of studies" or "poor quality reporting" problems generates a new dilemma: how to determine the reliability of the gathered knowledge. This is equivalent to the problem of determining the quality of the findings in an aggregation process including experimental studies that are not controlled and randomized clinical trials [Pino, J; 2004]. A possible solution or palliative for these problems is to build a reliability scale or "levels of evidence". This way, the person reading the results can take the necessary precautions concerning the generated knowledge.

In the following we describe the scale of levels of evidence that we are going to apply in this aggregation process:

- ***Level I****:* at this level of evidence we will be able to apply the experimental studies that have no reporting problems (which publish all the statistical parameters or original data to be able to estimate the parameters) provided that there is a normal distribution and homogeneity of variances in the behaviour of the phenomenon.

- ***Level II***: at this level of evidence we will be able to apply all the studies that apply level of evidence I, plus the experimental studies with slight reporting problems (mainly that do not publish the variance or standard deviation) or when there is not necessarily normality of distribution and homogeneity of variances in the phenomenon behaviour. In this case, the studies will be combined using non-parametric aggregation techniques.
- ***Level III***: at this level of evidence we will be able to apply all the studies that apply evidence level II, plus the experimental studies with serious reporting problems (unpublished variances or means) and/or studies that went through a process of "generalization" at the treatment level (generalization is grouping a set of similar treatments under the same name; for more details, see section step 3 of this chapter). In this case, the studies will be combined using non-statistical aggregation techniques. The findings will be general and will not be associated with a set confidence level. This means that the knowledge acquired from this technique will be less reliable.

Note importantly that, as mentioned above, if we consider it right to estimate level II evidence, there is nothing whatsoever to stop studies allocated to level I also being part of this new aggregation level. Following this criterion, Figure 9.1 illustrates how the better quality studies, which generate level I evidence, also output level II and level III evidence. The same applies to studies that generate level II knowledge, which can help to generate level III evidence.

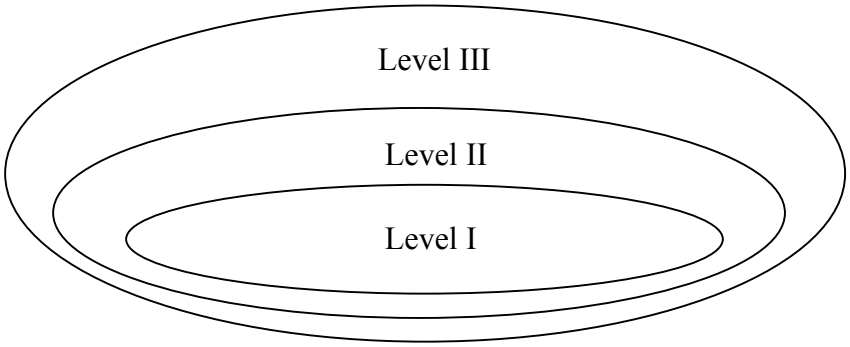

Figure 9.1: Venn diagram of the aggregation process

As Figure 9.1 shows, the goal of this procedure is to build pieces of knowledge ranging from the most to the least reliable and backed by the least to the most number of studies.

To get the above pieces of knowledge, the proposed aggregation process is divided into five steps: classify studies, analyse results, apply generalization strategies, aggregate studies and generate findings. Figure 9.2 below describes the sequence for executing the above steps:

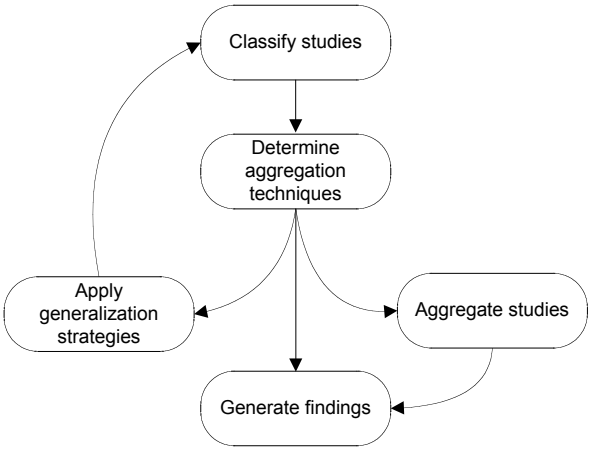

Figure 9.2: Sequence for executing the steps of the aggregation process

The objective of the steps specified in Figure 9.2 is described below:

- **Classify Studies:** The goal of classifying studies is to be able to group the different studies depending on their quality, the response variables they publish and the analysed treatment types.
- **Determine Aggregation Techniques:** The goal of this step is to identify the aggregation techniques to be applied depending on the number and quality of the detected studies. If no aggregation technique can be applied, an alternative step could be advisable: "Apply generalization strategies" or "Generate finding".
- **Apply Generalization Strategies:** The goal of the generalization strategy is to palliate the problems linked with the low number of replications. To do this, we have to look for common characteristics among the studies that can be used to group treatments and/or response variables at a higher level of abstraction (more general). Even though these groups with a higher level of abstraction are not genuine replications, they can, due to their similarity, be considered as conceptual replications, and, therefore, as studies that can be aggregated.
- **Aggregate Studies:** In this step, the different aggregation techniques will be applied to combine the results of the experimental studies. This is done based on the criteria and recommendations from the "Determine Aggregation Technique" step.
- **Generate Finding:** The goal of this step is to generate a report (as reliable as possible) from these pieces of knowledge, where the results analysis will start from the most reliable (gained by means of meta-analysis) to the least reliable knowledge (gained by means of alternative techniques). This way, if the results are compatible (all the evidence levels confirm that one treatment is better than another), we will have reached a more robust conclusion than we would have by applying the techniques separately. But if the results are not compatible, we should try to find out whether there are any as yet unidentified random variables at play or state the need to generate new experiments related to the subject.

The following sections detail each of the above steps.

9.2 DESCRIPTION OF THE AGGREGATION PROCESS STEPS

Step 1: Classify studies

To assure the reliability of the response estimated by the different aggregation techniques, one thing required of the included empirical studies is that they should meet a set of preconditions. Of these we consider it important to evaluate:

- Context Characteristics: This aspect is linked to two basic factors for applying the parametric techniques: distribution normality and homogeneity of variance.
- Reporting Completeness: This is a very important aspect, as no matter how well built the study is, if the report does not cover a minimum set of parameters, the aggregation techniques will not be able to be applied. The key parameters are: means (M), variances (V) and number of experimental subjects (N). It is also useful to identify if the report indicates whether or not the differences between the treatments is significant. Additionally, if no means are published, it can be helpful to know whether or not there were mean differences.
- Representativeness of the treatments and response variables: as there are not many study replications, this aggregation process proposes applying a generalization strategy (see "Apply Generalization Strategy" step) to put treatments that are not exactly the same but do have more in common than not in the same group. However, the differences between these generalized studies mean that statistical techniques are not applicable for estimation purposes. Briefly, treatment generalization limits the type of aggregation technique that can be used. The same applies to the representativeness of the generalized response variables.

Note that to be able to systematically categorize the different studies, it is necessary to make an additional decomposition depending on the response variables established in the research question or questions. Suppose, for example, that we want to find out which of two elicitation techniques, called "A" and "B", is better. To do this, the two variables defined for evaluation in the research question were the time it takes to develop the sessions and the number of requirements elicited. As these variables are not compatible with each other, we have to decompose the aggregation process into two groups "Technique A vs. Technique B using the session time response variable" and the "Technique A vs. Technique B using the number of requirements response variable". We will refer to this decomposition as the treatment-variable pair.
Table 9.1 below describes the key characteristics of the different categories and types of aggregation techniques that are recommended for use in each case.

Category	Characteristics of Studies	Technique Type
1	This category admits studies that have no biases and are similar in terms of their make-up and application domain, provided that there is a normal distribution and homogeneity of variance.	Parametric
2.1	This category admits studies with minor reporting defects (they do not publish standard deviations).	Non Parametric

Category	Characteristics of Studies	Technique Type
2.2	This category admits studies with moderate reporting defects (they do not publish standard deviations or means, but they do indicate whether there is a mean difference) and studies that generalize response variables.	Non Parametric
3	This category admits studies with serious reporting defects (they only express that a treatment is better than another without indicating the number of experimental subjects) and studies generalizing treatments.	Non Statistical

Table 9.1: Description of the categories of studies

Using the decision table described in Table 9.2 the studies are completely deterministically allocated to each category. This table indicates the minimum conditions that a study should meet to be put into a particular category.

Conditions	R1	R2	R3	R4
Context characteristics	The distribution is normal and there is homogeneity of variance	----	----	---
Report publishes	Means (Y), variances (s) and number of subjects (n)	Means (Y) and number of subjects (n)	That one treatment performs better than another and the number of subjects (n)	That one treatment performs better than another
Treatments and response variables	None were generalized	None were generalized	The response variables may have been generalized	The treatments and the response variables may have been generalized
Actions				
Place in Category	1	2.1	2.2	3

Table 9.2: Decision table for determining the category of studies

Example 9.1:

Problem definition:

Suppose that after searching and validating experiments comparing the use of an experimental treatment (E) and another control treatment (C) based on a single response variable called "Gain", we had 12 experimental studies. The key characteristics of all 12 studies are described in Table 9.3.

Id	Y^E	Y^C	n^E	n^C	S^E	S^C	Significant Differences	Observations
1	70	75	12	12	10	11	No	There is assumed to be a normal distribution and homogeneity of variance
2	105	90	8	8	15	14	Yes	There is assumed to be a normal distribution and homogeneity of variance
3	100	85	20	20	12	12	--	There is assumed to be a normal distribution and homogeneity of variance
4	95	100	4	4	----	----	--	----
5	130	75	20	20	----	----	Yes	----
6	100	60	24	24	----	----	Yes	----
7	95	80	50	50	----	----	No	----
8	----	----	50	50	----	----	Yes	The study indicates that the experimental treatment is better than the control treatment
9	----	----	50	50	----	----	--	The study indicates that the experimental treatment is better than the control treatment
10	----	----	50	50	----	----	--	The study indicates that the experimental treatment is better than the control treatment
11	100	60	24	24	----	----	--	Generalizes treatments
12	95	80	50	50	----	----	--	Generalizes treatments

Table 9.3: Summary of the experiment results

N.B.
Table 9.4 below describes the meaning of each of the columns in Table 9.3.

Column	Description
Id	Identifier of experimental study
Y^E / Y^C	Means of experimental and control treatments
n^E / n^C	Number of experimental subjects in experimental and control treatment
S^E / S^C	Standard deviations of the experimental and control treatments
Significant differences	Whether the experimental study indicates that the mean differences are significant for any hypothesis test
Observations	Additional information

Table 9.4: Description of columns in Table 9.3

Applying Step 1:

Table 9.5 below describes which category each experiment is placed into based on the recommendations in Decision Table 9.2.

Id	Category	Rule
1	1	R1
2	1	R1
3	1	R1
4	2.1	R2
5	2.1	R2
6	2.1	R2
7	2.1	R2
8	2.2	R3
9	2.2	R3
10	2.2	R3
11	3	R4
12	3	R4

Table 9.5: Description of study category

After we have placed each study into a category, let us define the treatment-variable pairs to organize how to account for the studies (in this case, there is only one):

1- "Experimental Treatment vs. Control Treatment" – "Gain"

Table 9.6 details all the articles for each category.

Treatment-Variable	Number of studies Category 1	Number of studies Category 2.1	Number of studies Category 2.2	Number of studies Category 3
"Experimental Treatment vs. Control Treatment" – "Gain"	3	4	3	2

Table 9.6: Number of studies in each category

Step 2: Analyse Results

To determine what techniques to use, we have to analyse how many studies are linked to each treatment-variable pair. This is because the precision of the techniques varies depending on how many studies there are [Lajeunesse, M & Forbes, M.; 2003]. For example, if there were more than 10 category 1 studies, it would be possible to apply parametric techniques (WMD and RR) only and end the process there, as the response reliability for these techniques with ten studies is very high [Lajeunesse, M & Forbes, M.; 2003].

Figure 9.3 below shows a flow chart describing how to select the aggregation techniques depending on how many studies have been identified for each category.

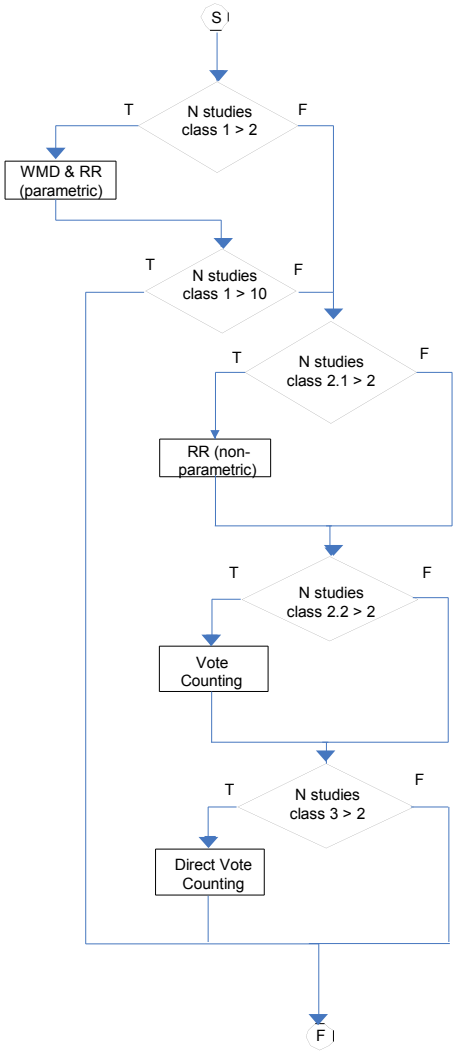

Figure 9.3: Aggregation technique categorization

N.B.
The N in the flow chart decision diamonds stands for the number of experimental studies.

The decisions shown in Figure 9.3 should be interpreted as follows:

- ✓ If there are more than 10 category 1 studies, use WMD to aggregate those studies as in standard processes and end the process.
- ✓ If there are less than 10 category 1 studies then,
  - If there are two or more category 1 studies, preferably use WMD and parametric RR for aggregation.
  - If there are category 2.1 studies, add to the category 1 studies and preferably use non-parametric RR for aggregation.
  - If there are category 2.2 studies, add to the category 1 and category 2.1 studies and preferably use vote counting for aggregation.
  - Finally, if there are category 3 studies, add all four categories of studies together and preferably use direct vote counting for aggregation.

N.B.
At this point the aggregation process could be concluded if there is considered to be a shortage of studies and it not feasible to apply generalization to try to salvage the aggregation process by searching for new alternatives for combining studies.

Example 9.2:

Applying Step 2:

Going back to example 9.1, Table 9.7 below again shows the number of studies per category (a copy of Table 9.5). These data and the decision rules from Figure 9.3 are used to determine what aggregation techniques to apply to each study.

Treatment-Variable	Number of category 1 studies	Number of category 2.1 studies	Number of category 2.2 studies	Number of category 3 studies
"Experimental Treatment vs. Control Treatment" – "Gain"	3	4	3	2

Table 9.7: Number of studies allocated to each category

As Table 9.7 shows, there are less than 10 category 1 studies. This means that all the aggregation techniques should be applied. Briefly, Table 9.8 states which aggregation techniques will be applied to each study. To do this, we used the decision rules described in Figure 9.3.

Id	WMD	Parametric RR	Non-Parametric RR	Vote Counting	Direct Vote Counting
1	X	X	X	X	X
2	X	X	X	X	X
3	X	X	X	X	X
4			X	X	X
5			X	X	X
6			X	X	X
7			X	X	X
8				X	X
9				X	X
10				X	X
11					X
12					X

Table 9.8: Aggregation techniques to be applied

Step 3: Apply generalization strategy

This step has two goals: a) to reduce problems of bias by means of an analysis interpreting the treatments and response variables and b) to palliate problems of there not being enough replications by generalizing the treatments and response variables. In the following we describe what each task entails:

- Interpretation:
  Interpretation involves identifying whether two identical treatments or response variables might be being considered different due to a lexical problem. This can often happen due to translation problems or the omission of part of the treatment name by the paper authors.
  The interpretation process involves reading in the detail the section of the article describing what the treatments and response variables involve and evaluating whether, even though they have different names, they are essentially the same.
  Note that the process of interpretation has no drawbacks regarding results combination, and its purpose is to reduce publishing bias in order to increase the evidence level.
- Generalization:
  The goal of generalization is to show up common aspects at a higher level of abstraction between two treatments or response variables. Commonly, the techniques or variables used in software engineering can be clustered depending on a particular feature. When there are not a great many studies in an aggregation process and we have identified a set of studies that cannot be aggregated because they do not directly answer the research question, we can try to enact a generalization process.
  The process of generalization involves reading in detail the section of the article describing what the treatments and response variables involve and evaluating whether there is any high-level concept under which they can be grouped as a more general set.

N.B.
Note that when the generalization step is applied, we have to go back to step 1 and evaluate and classify studies again.

Example 9.3:
Applying Step 3:

In the following we describe two examples. One is linked to interpretation and the other to generalization. The examples are not related to exercises 9.1 and 9.2 because they only reported the numerical values of the statistical parameters.

Interpretation:
Interviewing is a typical example of imprecise treatment referencing. A structured interview is often referenced as a conventional interview or simply interview. In these cases, if we do not evaluate what the technique in the article involves, we will not be able to find out exactly whether it is a structured or

open interview, as they can both be considered conventional and they are both interviews.

Generalization:
Suppose that we were trying to find out whether or not the "C++" language is better than its predecessor "C", and we had only two studies comparing these two languages directly. The chances of the aggregation process generating reliable results with this evidence are very scant. But, what, if apart from these two studies, we had other studies comparing "Delphi" and "Pascal"? As "C++" and "Delphi" are object oriented and "C" and "Pascal" are structured, we can hypothesize that by "lumping together" or generalizing "C++" and "Delphi" (as well as "C" and "Pascal"), we will be able to reach a more reliable conclusion because there are more studies. Obviously, these findings do not answer the question "is C++ better than C?", but they do respond to another very similar question that generates knowledge about the first.

Step 4: Aggregate Studies

In this step we will apply the different aggregation techniques for combining the results of the experimental studies. This will based on the criteria and recommendations made in step 2 "Determine Aggregation Techniques". Apart from estimating the effect index for the cases where WMD or RR (both versions) are applied, we also have to use forest plots to evaluate the heterogeneity between the experiments.

Example 9.4:

Applying Step 4:

Going back to examples 9.1 and 9.2, Table 9.9 below again shows the table of aggregation techniques to be applied (it is a copy of Table 9.8).

Id	WMD	Parametric RR	Non-Parametric RR	Vote Counting	Direct Vote Counting
1	X	X	X	X	X
2	X	X	X	X	X
3	X	X	X	X	X
4			X	X	X
5			X	X	X
6			X	X	X
7			X	X	X
8				X	X
9				X	X
10				X	X
11					X
12					X

Table 9.9: Aggregation techniques to be applied

The following sections describe the application of the techniques recommended in Table 9.9

Applying parametric techniques

According to the recommendations in Table 9.8 experiments 1, 2 and 3 can be aggregated using parametric techniques: weighted mean differences and parametric response ratio.

Table 9.10 below describes the values of the above studies.

Id	Y^E	Y^C	n^E	n^C	S^E	S^C	Observation
1	70	75	12	12	10	11	----
2	105	90	8	8	15	14	----
3	100	85	20	20	12	12	----

Table 9.10: Values reported in studies

Estimating the Effect Size (WMD)

Table 9.11 shows the results.

Study	Effect	Upper Bound	Lower Bound	Weight
1	-1.270	0.351	-0.459	32.77%
2	-0.059	2.014	0.977	20.03%
3	0.550	1.901	1.225	47.20%
Overall	0.624	0.160	1.080	

Table 9.11: Results

Figure 9.4 graphically shows the results reported in Table 9.11.

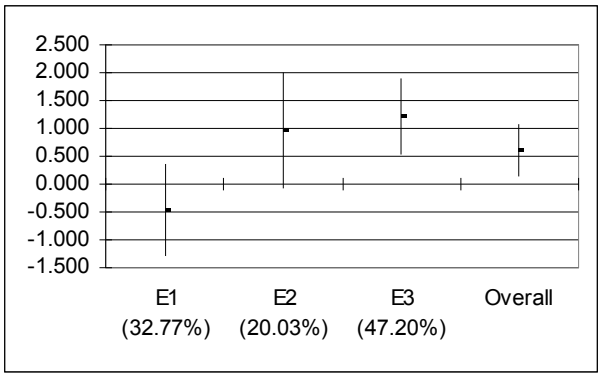

Figure 9.4: Results

We find that there is homogeneity between the results of the studies described in Figure 9.4, as the confidence intervals of the different studies overlap with the confidence interval of the overall effect.

Estimating the Parametric RR

Table 9.12 shows the results.

Id	RR	Upper Bound	Lower Bound	Weight
1	0.933	0.831	1.048	27.31%
2	1.167	1.008	1.351	17.11%
3	1.176	1.085	1.276	55.57%
Overall	1.102	1.038	1.171	

Table 9.12: Results

Figure 9.5 graphically shows the results reported in Table 9.12.

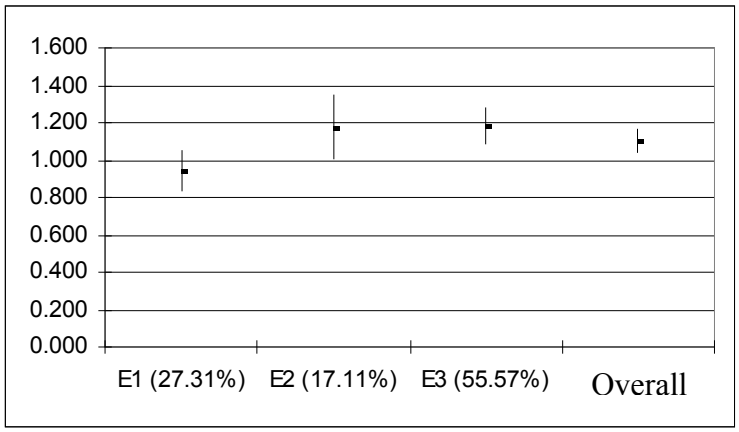

Figure 9.5: Results

We find that there is homogeneity between the results of the studies described in Figure 9.5, as the confidence intervals of the different studies overlap with the confidence interval of the overall effect.

Applying Non-Parametric Techniques

Non-Parametric Response Ratio

Following the recommendations in Table 4.8, experiments 1, 2, 3, 4, 5, 6 and 7 can be aggregated using the Non-Parametric Response Ratio.

Table 9.13 below describes the results of the studies that can be aggregated using this technique.

Id	Y^E	Y^C	n^E	n^C	Observation
1	70	75	12	12	----
2	105	90	8	8	----
3	100	85	20	20	----
4	95	100	4	4	----
5	130	75	20	20	----
6	100	60	24	24	----
7	95	80	50	50	----

Table 9.13: Values reported in the studies

Table 9.14 shows the results.

Study	Effect	Upper Bound	Lower Bound	Weight

Table 9.14: Results

Figure 9.6 graphically shows the results reported in Table 9.14:

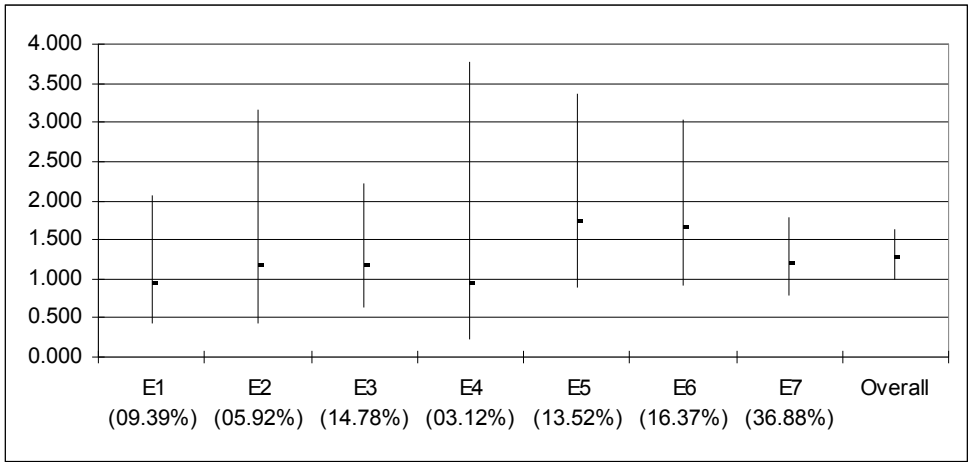

Figure 9.6: Results

We find that there is homogeneity between the results of the studies described in Figure 9.6, as the confidence intervals of the different studies overlap with the confidence interval of the overall effect.

Vote Counting

Following the recommendations in Table 4.8, experiments 1, 2, 3, 4, 5, 6, 7, 8, 9 and 10 can be aggregated using vote counting.

Table 9.15 below describes the results of the studies that can be aggregated by this technique.

8	----	----	50	50	The study indicates that the experimental treatment is better than the control treatment
9	----	----	50	50	The study indicates that the experimental treatment is better than the control treatment
10	----	----	50	50	The study indicates that the experimental treatment is better than the control treatment

Table 9.15: Values reported in the studies

Table 9.16 shows how votes are assigned to each study.

Id	Vote
1	0
2	1
3	1
4	0
5	1
6	1
7	1
8	1
9	1
10	1

Table 9.16: Voting

Table 9.17 below shows the estimated effect.

Effect	Upper Bound	Lower Bound
0.35	0.26	0.44

Table 9.17: Results

Figure 9.7 graphically shows the results reported in Table 9.17.

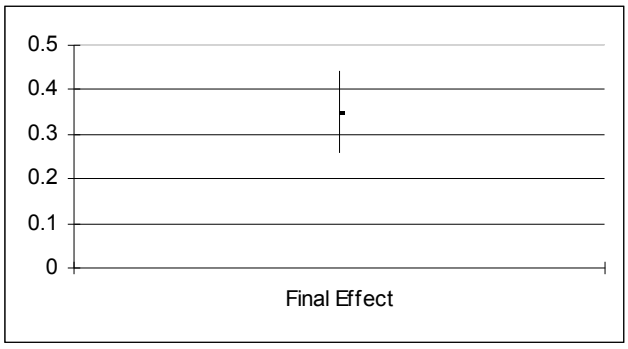

Figure 9.7: Results

Direct Vote Counting

Following the recommendations in Table 4.8, experiments 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11 and 12 can be aggregated using Direct Vote Counting.

Table 9.18 below describes the results of the studies that can be aggregated using this technique.

Id	Y^E	Y^C	Sig. Diff.	Observation
1	70	75	No	----
2	105	90	Yes	----
3	100	85	--	----
4	95	100	--	----
5	130	75	Yes	----
6	100	60	Yes	----
7	95	80	No	----
8	----	----	Yes	The study indicates that the experimental treatment is better than the control treatment
9	----	----	--	The study indicates that the experimental treatment is better than the control treatment
10	----	----	--	The study indicates that the experimental treatment is better than the control treatment
11	100	60	--	The study indicates that the experimental treatment is better than the control treatment
12	95	80	--	The study indicates that the experimental treatment is better than the control treatment

Table 9.18: Values reported in studies

Before voting takes place, let us set the cut-off value at a minimum of 51% votes in favour with 30% of significant votes.

Table 9.1 below describes the voting for each study.

Id	Vote
1	-
2	++
3	-
4	++
6	++
7	+
8	++
9	+
10	+
11	+
12	+

Table 9.19: Experiment results

Table 9.20 shows the counted votes.

Vote category	Number of votes	%
++	4	33.3
+	6	50
O	0	0
-	2	16.6
--	0	0

Table 9.20: Results

Figure 9.8 graphically shows the results reported in Table 9.20.

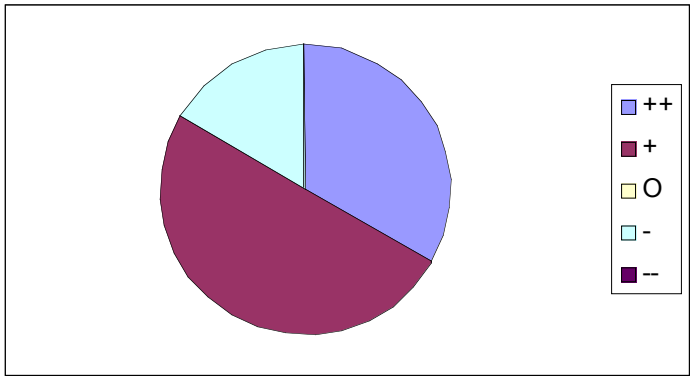

Figure 9.8: Results

Step 5: Generate Conclusion

In this step we describe a strategy for putting together the final report describing the knowledge gathered during the process. In this respect, it is important to note that when generating conclusions: *"Statistics are an aid but not a substitute for the expert's judgement. All these numbers aspire to add to the discussion [Cochrane; 2008]"*.

Ideally, the discussion of and conclusions from the findings of an aggregation process should take a broad perspective targeting a wide range of potential readers. In this respect, we consider that the conclusions should state:

- The level of quality and number of studies covered
- The level of reliability of the aggregation techniques used
- How broad and significant the estimated effects are
- How much bearing each of the studies has on the final conclusion
- How consistent the between-study effects are
- How consistent the effects between the evidence levels are

N.B.
When more than one evidence level is generated, the possibilities are:

- All the evidence levels provide compatible results (e.g. indicate that the experimental treatment is better than the control treatment). In this case, all the evidence could be said to confirm the same hypothesis,
- The results between the different evidence levels are incompatible (e.g. one level indicates that the experimental treatment is better than the control treatment and another level indicates the opposite). If this is the case, the less reliable studies will have to be analysed in more detail to try to find out whether there are unidentified independent variables that are affecting the results and, if necessary, divide the aggregation process into two or more groups.

The above knowledge should be put together in a final report, which should contain the following sections:

1. "Pieces of knowledge gathered":
   This describes the results for the different evidence levels, detailing the above aspects.
2. "Possible research lines":
   This describes the knowledge for which there is no firm evidence or is considered to need further investigation.

Example 9.5:

Applying Step 5:

Below we present the findings after having aggregated the studies from example 9.4.

Section I – Acquired pieces of knowledge

1. The aggregation process combined 12 studies with the following breakdown:

Evidence level	Number of studies	Total number of experimental subjects
I	3	80
II	10	576
III	12	724

Table 9.21: Studies by evidence levels

2. Analysis of estimated effects:

Evidence Level	Applied Technique	Findings
I	WMD	➢ The final estimated effect is **0.624**, which, according to the results interpretation tables, is a **medium effect.** This means that the experimental treatment is better than the control treatment, but, in principle, not by very much. ➢ The fact that the lower bound of the confidence interval is above zero lends strength to the hypothesis that the experimental treatment is better than the control treatment.
I	Parametric RR	➢ The final estimated ratio is **1.102**, implying that the experimental treatment is **10% better** than the control treatment. ➢ The fact that the lower bound of the confidence interval is above one lends strength to the hypothesis that the experimental treatment is better than the control treatment.
II	Non-parametric RR	➢ The final estimated ratio is **1.279** which implies that the experimental treatment is almost **30% better** than the control treatment. ➢ As the confidence interval of the final effect overlaps with the confidence interval of each study, there can be said to be homogeneity.
II	Vote Counting	➢ The final estimated effect is **0.35**, which, according to the results interpretation tables, implies a **medium effect**. This means that the experimental treatment is better than the control treatment, but, in principle, not by very much. ➢ As the confidence interval of the final effect overlaps with the confidence intervals of each study, they can be said to be homogeneous.
III	Direct Vote Counting	➢ **83.3%** of studies (votes ++ and +) indicate that the experimental treatment is better than the control treatment ➢ **33.3%** of the studies (votes ++) indicate that there is significant evidence in favour of the experimental treatment.

Table 9.22: Estimated effects by evidence levels

3. Looking at the weights of the different studies, study "3" can be said to have most influence on level-I evidence and study "3" and study "7" are

the most influential level-II studies. Note that both studies (the largest in terms of experimental subjects) have very similar effects.

4. The results of the different evidence levels can be said to be absolutely consistent. All the levels indicate that the experimental treatment is better than the control treatment. The inclusion of studies aggregated by less precise techniques can extend the empirical evidence.
5. The more studies there are in the aggregation process, the more marked the differences in favour of the experimental treatment are.

Section II – Possible Research Lines

- It would be important to have more high quality empirical studies to be able to corroborate the trends that appear to indicate that the more studies there are, the greater the improvement of the experimental treatment compared with the control treatment.

10 References

Agarwal, R.; Tanniru, M.; 1990; *Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation;* Journal of Management Information System, M.E. Sharpe; Vol. 7 N. 1

Banker and Keremer; 1989; *Scale economies in new software development.* IEEE Transactions on Software Engineering. (15): 10, pp. 1199-1205.

Borenstein, M.; Hedges, L; Rothstein, H.; 2007; Meta-Analysis Fixed Effect vs. random effect; WWW.Meta-Analysis.com

Basili, V.; Green, S.; Laitenberger, O.; Lanubile, F.; Shull, F.; Sorumgaard, S.; Zelkowitz, M.; 1996; Packaging researcher experience to assist replication of experiments ISERN Meeting

Burton, A., Shadbolt, N., Hedgecock, A. and Rugg, G.; 1988; *A Formal Evaluation of Knowledge Elicitation Techniques for Expert Systems: Domain 1*. Proceedings of Expert Systems '87 on Research and Development in Expert Systems IV**.** pp. 136-145.

Burton, A., Shadbolt, N., Rugg, G. and Hedgecock, A.; 1990. *The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise*. Knowledge Acquisition 2(2): 167-178.

Cochrane; 2008; *Curso Avanzado de Revisiones Sistemáticas*; www.cochrane.es/?q=es/node/198

Conover W. *Practical Nonparametric Statistics*. 2nd ed. New York: John Wiley & Sons; 1980.

Crandall Klein, B.; 1989. *A Comparative Study Of Think-Aloud And Critical Decision Knowledge Elicitation Method*. SIGAR Newsletter, April 1989, Number 108, Knowledge Acquisition Special Issue, pp 144-146.

Davis, A.; Dieste O.; Hickey, A.; Juristo, N.; Moreno, A.; 2006; *Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review*; 14th IEEE International Requirements Engineering Conference (RE'06) pp. 179-188

DerSimonian R. and Laird N.; *Meta-analysis in clinical trials*; Control Clin Trials 1986; 7: 177-88.

Dixon-Woods, M.; Agarwal, S.; Jones, D.; Young, B., and Sutton, A.; 2005; *Synthesising qualitative and quantitative evidence: a review of possible methods*. Journal of Health Services Research and Policy. ; 10(1):45-53B(9).

Dyba, T., Aricholm, E.; Sjoberg, D.; Hannay J.; Shull, F.; 2007; Are two heads better than one? On the effectiveness of pair programming. IEEE Software;12-15.

Epidat, 2008; programa de libre distribución desarrollado por instituciones públicas y dirigido a epidemiólogos y otros profesionales de la salud; www.paho.org/spanish/sha/epidat.htm

Evidence-Based Medicine Working Group; 1992; Evidence-based medicine. A new approach to teaching the practice of medicine. JAMA, 268(17), 2420-2425.

Fairbank L, O'Meara S, Renfrew MJ, Woodridge M, Sowden AJ, Lister-Sharp D.; 2000; *A systematic review to evaluate the effectiveness of interventions to promote the initiation of breastfeeding*. Health Technology Assessment; 4: 1–171

Fernandez, E.;2007; Aggregation process with multiple evidence level for experimental studies in software engineering; 2nd International Doctoral Symposium on Empirical Software Engineering (IDoESE); 75-81

García, R.; 2004; Inferencia Estadística y Diseño de Experimentos; eudeba; Buenos Aires Argentina

Gardner M., Altman D.; *Statistics with confidence. Confidence intervals and statistical guidelines*. London: BMJ; 1992.

Glass, G; 1976; Primary, secondary, and meta-analysis of research. Educational Researcher 5: 3-8

Glass, G; 2000; Meta-Analysis at 25. http//glass.ed.asu.edu/gene/papers/meta25.html

Goodman C.; 1996*; Literature Searching and Evidence Interpretation for Assessing Health Care Practices;* SBU; Stockholm.

Grimán Padua; 2007; Proposal of a review process of empirical studies in software engineering; International Doctoral Symposium on Empirical Software Engineering (IDoESE). 25-32

Guerra Romero, L.; 1996; La medicina basada en evidencia: un intento de acercar la ciencia al arte de la práctica médica; Plan Nacional sobre el Sida. Ministerio de Sanidad y Consumo, Madrid, España; Med Clin (Barc) 1996; 107:377-382

Gurevitch, J. and L.V. Hedges, 1993. Meta-analysis: Combining the results of independent experiments. Pages 378-398 in S.M. Scheiner and J. Gurevitch, editors. Design and Analysis of Ecological Experiments. Chapman and Hall, New York.

Gurevitch, J. and Hedges, L.; 2001; *Meta-analysis: Combining results of independent experiments.* Design and Analysis of Ecological Experiments (eds S.M. Scheiner and J. Gurevitch), pp. 347–369. Oxford University Press, Oxford.

Hedges, L.; Gurevitch, J.; Curtis, P.; 1999; The Meta-Analysis of Response Ratio in Experimental Ecology; The Ecological Society of America

Hedges, L.; Olkin, I.; 1985; *Statistical methods for meta-analysis*. Academic Press.

Hu, Q.; 1997; *Evaluating Alternative Software Production Function*. IEEE Transactions on Software Engineering. (23): 6, pp. 379-387.

Jørgensen, M.; 2004; *A Review of Studies on Expert Estimation of Software Development Effort*. Journal of Systems and Software. (70): 1-2, pp. 37-60.

Juristo, N.; Moreno, A. M., and Vegas, S.; 2004; *Reviewing 25 Years of Testing Technique Experiments*. Journal of Empirical Software Engineering; 9(1 - 2):7-44.

Kitchenham, B. A.; 2004; *Procedures for performing systematic reviews*. Keele University; TR/SE-0401. Keele University Technical Report.

Lajeunesse, M & Forbes, M.; 2003; *Variable reporting and quantitative reviews: a comparison of three meta-analytical techniques*. Ecology Letters, 6: 448-454.

Meta-Analysis; 2008; WWW.Meta-Analysis.com

Miguez, E. & Bollero, G; 2005; Review of Corn Yield Response under winter cover cropping systems using Meta-Analytic Methods; Crop Science Society of America

Miller, J.; 2000; *Applying Meta-analytical Procedures to Software Engineering Experiments*. Journal of Systems and Software. (54): 1, pp. 29-39.

Mohagheghi, P., & Conradi , R.; 2004; *Vote-Counting for Combining Quantitative Evidence from Empirical Studies - An Example*. Proceedings of the International Symposium on Empirical Software Engineering (ISESE'04) .

Molinero, L.; 2006; Meta-Análisis; Asociación de la Sociedad Española de Hipertensión

Neyman J, Pearson E. *On the problem of the most efficient tests of statistical hypotheses*. *Philosophical Trans of the Royal Society of London A* 1933; 231: 289-337.

Noblit, G. W., & Hare, R. D.; 1988; Meta-Ethnography: Synthesising Qualitative Studies. Newbury Park, CA: Sage.

Pickard, L. M.; Kitchenham, B. A., and Jones, P. W.; 1998; *Combining empirical results in software engineering. Information and Software Technology*.; 40(14):811-821.

Pillemer, D. and Light, R.; 1980; *Synthesizing outcomes: How to use research evidence from many studies*. Harvard Educational Review.

Primo, J.; 2004; *Niveles de evidencia y grados de recomendación*. Ponencia presentada en el Symposium "Gestión del conocimiento y su aplicación en la Enfermedad Inflamatoria Crónica Intestinal", organizado por GETECCU (Grupo Español de Trabajo en Enfermedad de Crohn y Colitis Ulcerosa).

Shekelle, P.; Maglione, M.; Morton, S.; 2003; Judging What to Do About Ephedra; http://rand.org/publications/randreview/issues/spring2003/evidence.htm

Shull, F.; Carver, J.; Travassos, G. H.; Maldonado, J. C.; Conradi, R., and Basili, V. R.; 2003; *Replicated Studies: Building a Body of Knowledge about Software Reading Techniques.* Lecture Notes on Empirical Software Engineering. Chapter 2, pp. 39-84. World Scientific.

Straus, S. E. ; Richardson, W. S.; Glasziou, P., and Haynes, R. B.; 2005; *Evidence*-Based Medicine. Churchill Livingstone.

Takkouche B, Cadarso-Suarez C, Spiegelman D. *Evaluation of old and new tests of heterogeneity in epidemiologic meta-analysis*. Am J Epidemiol 1999; 150: 206-15.

Thalheimer W. and Cook S.; 2002; How to calculate effect sizes from published research: A simplified methodology. A Work-Learning Research Publication.

Wohlin, C., Petersson, H., & Aurum, A.; 2003; Combining data from reading experiments in software inspections: a feasibility study. (pp. 85-132). World Scientific Publishing Co., Inc.

Woody, J.; Will, R.; Blanton, J.; 1996; *Enhancing Knowledge Elicitation using the Cognitive Interview;* Expert system with application; Vol. 10 N. 1

Worn, B.; Barbier, E.; Beaumont, N.; Duffy, J.; Folke, C; Halpern, B.; Jackson, J.; Lotze, H.; Micheli, F.; Palumbi, S.; Sala, E.; Selkoe, K.; Stachowics, J.; Watson, R; 2007; Supporting Online Material: Impacts of biodiversity loss on ocean ecosystem services.

Yin, R. K. and Heald, K. A.; 1975; *Using the Case Survey Method to Analyze Policy Studies*. Administrative Science Quarterly; 20(3):371-381.

Appendix A

In this section we describe the data collection forms of the studies that were part of the aggregation process.

Empirical study 1:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>A Formal Evaluation of Knowledge Elicitation Techniques for Expert Systems: Domain 1</td></tr>
<tr><td>2</td><td>Reference</td><td>Burton, A., Shadbolt, N., Hedgecock, A. & Rugg, G. 1988. A Formal Evaluation of Knowledge Elicitation Techniques for Expert Systems: Domain 1. Proceedings of Expert Systems '87 on Research and Development in Expert Systems IV. pp 136-145.</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Requirements elicitation for “rock identification”</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Interviewing
➢ Protocol analysis
➢ Card sorting
➢ Laddering</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>The experts working on the project were 3rd-year geology degree students.
They worked with 32 experts divided into 4 8-person groups. Each technique was tested by 16 experts:
n = 16.</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Elicitation session time
2- Elicitation session transcription time
3- Total time (effort)
4- Number of elicited rules
5- Number of clauses (gain)
6- Rule completeness</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Variable 3 is equal to variables 1 plus 2. Therefore, they will not be taken into account here.
Variable 4 is a less accurate way of identifying the gain. Therefore, it will not be taken into account.
Variables 3, 5 and 6 apply in this analysis.</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th></th><th>Inter-viewing</th><th>Protocol analysis</th><th></th><th>Ladder-ing</th><th>Card sorting</th><th></th></tr>
<tr><td>Variable</td><td>μ</td><td>μ</td><td>F</td><td>μ</td><td>μ</td><td>F</td></tr>
<tr><td>Effort (minutes)</td><td>80.8</td><td>110.3</td><td>4.2</td><td>79.8</td><td>67.0</td><td>4.2</td></tr>
<tr><td>Gain</td><td>94.4</td><td>75.8</td><td></td><td>101.4</td><td>63.4</td><td></td></tr>
<tr><td>Complete-ness</td><td>27.9</td><td>7.9</td><td>39.3</td><td>28.1</td><td>30</td><td>39.3</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 2:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>Laddering: Technique and Tool in Knowledge Acquisition</td></tr>
<tr><td>2</td><td>Reference</td><td>Corbridge, C., Rugg, G., Major, P., Shadbolt, N. & Burton, A. 1994. Laddering: Technical and Tool in Knowledge Acquisition. Department of Psychology, University of Nottingham; Nottingham NG7 2RD.</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Requirements elicitation for “abdominal conditions” diagnosis.</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Interviewing
➢ Protocol analysis
➢ Card sorting
➢ Laddering</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>The experts working on this study were final-year medical students. On privacy grounds, real patients were not used. The students played the role of patients.
They worked with 32 experts divided into 8 4-member groups.
n = 16.</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Time (effort)
2- Number of clauses (gain)</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>All the variables apply in this analysis.</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th></th><th colspan="2">Inter-viewing</th><th colspan="2">Protocol analysis</th><th colspan="2">Laddering</th><th colspan="2">Card sorting</th></tr>
<tr><td>Variable</td><td>μ</td><td>S</td><td>μ</td><td>S</td><td>μ</td><td>S</td><td>μ</td><td>S</td></tr>
<tr><td>Effort (minutes)</td><td>39.5</td><td>14,3</td><td>26.75</td><td>5,14</td><td>40.75</td><td>16</td><td>29.75</td><td>13,6</td></tr>
<tr><td>Gain</td><td>274</td><td>102</td><td>145</td><td>74</td><td>521.4</td><td>420</td><td>144</td><td>52</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 3:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise</td></tr>
<tr><td>2</td><td>Reference</td><td>Burton, A., Shadbolt, N., Rugg, G. & Hedgecock, A. 1990. The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise. Knowledge Acquisition 2(2): 167-178.</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Requirements elicitation for evaluating the identification of “rock artefacts”</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Interviewing
➢ Protocol analysis
➢ Card sorting
➢ Laddering</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>They worked with real geology experts. Four Experts applied each technique
n = 4.</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Time (effort)
2- Number of clauses (gain)</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>All the variables apply to this analysis.</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th></th><th>Interviewing</th><th>Protocol analysis</th><th>Laddering</th><th>Card sorting</th></tr>
<tr><td>Variable</td><td>μ</td><td>μ</td><td>μ</td><td>M</td></tr>
<tr><td>Effort (minutes)</td><td>217</td><td>351</td><td>98</td><td>145</td></tr>
<tr><td>Gain</td><td>270</td><td>269</td><td>123</td><td>188</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 4:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise</td></tr>
<tr><td>2</td><td>Reference</td><td>Burton, A., Shadbolt, N., Rugg, G. & Hedgecock, A. 1990. The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise. Knowledge Acquisition 2(2): 167-178.</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Requirements elicitation to evaluate the identification of “pottery shards”</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Interviewing
➢ Protocol analysis
➢ Card sorting
➢ Laddering</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>They worked with real pottery experts. Four experts applied each technique
n = 4.</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Time (effort)
2- Number of clauses (gain)</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>All the variables apply to this analysis.</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th></th><th>Interviewing</th><th>Protocol analysis</th><th>Laddering</th><th>Card sorting</th></tr>
<tr><td>Variable</td><td>μ</td><td>μ</td><td>μ</td><td>μ</td></tr>
<tr><td>Effort (minutes)</td><td>240</td><td>176</td><td>145</td><td>177</td></tr>
<tr><td>Gain</td><td>317</td><td>184</td><td>216</td><td>278</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 5:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>Comparing Knowledge Elicitation Techniques: A Case Study</td></tr>
<tr><td>2</td><td>Reference</td><td>Schweickert, R., Burton, A., Taylor, N., Corlett, E., Shadbolt, N., Rugg, G. & Hedgecock, A.; 1987. Comparing Knowledge Elicitation Techniques: A Case Study. Artificial Intelligence Review (1): 245-253.</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Eliciting requirements for the choice of “special lighting” for surface analysis</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Interviewing
➢ Twenty Questions
➢ Card sorting</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>They worked with one real lighting expert.
Two sessions were held for each technique, and all the sessions involved the same expert.
n = 2</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Number of rules
2- Percentage of implemented rules</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Variable 1 is a less precise way of evaluating “gain”
Variable 2 is an alternative to evaluating rule completeness.</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th></th><th>Interviewing</th><th>Twenty Questions</th><th>Card Sorting</th></tr>
<tr><td>Variable</td><td>μ</td><td>μ</td><td>μ</td></tr>
<tr><td>Gain</td><td>61</td><td>50</td><td>10</td></tr>
<tr><td>Completeness</td><td>59 %</td><td>48 %</td><td>56 %</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Questionable
It does not describe all the techniques defined in the “question under assessment”
The comparisons do not directly match the “question under assessment”
The gain is inferred from rules and not clauses
The completeness is inferred from the number of implemented rules and not the number of rules defined by a senior expert.</td></tr>
</table>

Empirical study 6:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>A Comparative Study Of Think-Aloud and Critical Decision Knowledge Elicitation Method</td></tr>
<tr><td>2</td><td>Reference</td><td>Crandall Klein, B. and Associates; 1989. A Comparative Study Of Think-Aloud And Critical Decision Knowledge Elicitation Method. SIGAR Newsletter, April 1989, Number 108, Knowledge Acquisition Special Issue, pp 144-146.</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Requirements elicitation for “Extinguishing a Fire”</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Think-Aloud
➢ Critical Decision Method</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>They worked with fire-fighters with 20 years’ experience and 11 years’ experience as team leaders.
The experts were divided into two groups, and each group participated in the use of one technique.
Fire drills were carried out.
n = 10</td></tr>
<tr><td>7</td><td>Response variables</td><td>Quantity of information indicating which technique was the best without giving real values.</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Variable 1 is a less precise alternative for evaluating “gain”</td></tr>
<tr><td>9</td><td>Results</td><td><table><tr><th></th><th>Think-Aloud</th><th>Critical Decision Method</th></tr><tr><td>Gain</td><td colspan="2">The “Critical Decision Method” technique gathers more information than the “Think-Aloud” technique</td></tr></table></td></tr>
<tr><td>10</td><td>Study rating</td><td>Questionable
It does not describe exactly the same techniques as were defined in the “question under assessment”
The gain is inferred qualitatively</td></tr>
</table>

Empirical study 7:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>A comparison of five elicitation techniques for elicitation of attributes of low involvement products</td></tr>
<tr><td>2</td><td>Reference</td><td>Bech-Larsen, T., Nielsen, N.,. 1997. A comparison of five elicitation techniques for elicitation of attributes of low involvement products. The MAPP Centre, The Aarhus School of Business, Haslegaardsvej 10, DK-8210 Aarhus V, Denmark Received 12 December 1997; received in revised form 17 July 1998; accepted 27 July 1998.</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Selection of oils in a supermarket</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Triadic sorting
➢ Free sorting
➢ Direct sorting
➢ Ranking
➢ Picking from an attribute list</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>They interviewed a set of supermarket customers
n = 30</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Attribute abstraction level
2- Attribute importance
3- Importance of the first five attributes
4- Ability to discriminate alternative products
5- Ability to discriminate products
6- Number of elicited attributes
8- Attribute predictability</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Variable 6 is considered to be the best response variable for this problem. The other variables will not be taken into account.</td></tr>
<tr><td>9</td><td>Results</td><td><table><tr><th>Technique</th><th>Number of attributes</th></tr><tr><td>Triadic sorting</td><td>6.46</td></tr><tr><td>Free sorting</td><td>7.70</td></tr><tr><td>Direct sorting</td><td>8.60</td></tr><tr><td>Ranking</td><td>9.53</td></tr><tr><td>Picking from an attribute list</td><td>9.83</td></tr></table></td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 8:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques</td></tr>
<tr><td>2</td><td>Reference</td><td>Breivik, E., Supphellen, M. 2001. Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques. Norwegian School of Economics and Business Administration, Institute of Strategy and Management, Breiviksveien 40, 5045 Bergen, Norway Received 24 May 2001; accepted 5 November 2001</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Selection of attributes characterizing a product</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Direct elicitation
➢ Rank ordering elicitation.
➢ Ideal description.</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>Consumers asked over the telephone about the distinctive features of a car.</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Number of attributes
2- Importance of attributes</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Variable 1 is considered the best response variable for this problem. The other variables will not be taken into account.</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th>Technique</th><th>Number of attributes</th><th>n</th><th>s</th></tr>
<tr><td>Direct elicitation</td><td>4.49</td><td>43</td><td>1.40</td></tr>
<tr><td>Rank ordering elicitation</td><td>4.32</td><td>39</td><td>2.00</td></tr>
<tr><td>Ideal description</td><td>4.33</td><td>37</td><td>1.67</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 9:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques</td></tr>
<tr><td>2</td><td>Reference</td><td>Breivik, E., Supphellen, M. 2001. Elicitation of product attributes in an evaluation context: A comparison of three elicitation techniques. Norwegian School of Economics and Business Administration, Institute of Strategy and Management, Breiviksveien 40, 5045 Bergen, Norway Received 24 May 2001; accepted 5 November 2001</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Selection of attributes characterizing a product</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Direct elicitation
➢ Rank ordering elicitation
➢ Ideal description</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>Consumers asked over the telephone about the aspects they rate on restaurant menus</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Number of attributes
2- Importance of attributes</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Variable 1 is considered the best response variable for this problem. The other variables will not be taken into account.</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th>Technique</th><th>Number of attributes</th><th>n</th><th>s</th></tr>
<tr><td>Direct elicitation</td><td>4.95</td><td>43</td><td>1.60</td></tr>
<tr><td>Rank ordering elicitation</td><td>4.85</td><td>39</td><td>1.83</td></tr>
<tr><td>Ideal description</td><td>4.08</td><td>37</td><td>1.44</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 10:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>An empirical Investigation of User Requirements Elicitation: Comparing the Effectiveness of Prompting Techniques</td></tr>
<tr><td>2</td><td>Reference</td><td>Browne, G.; Rogich, M.; An Empirical Investigation of User Requirements Elicitation: Comparing the Effectiveness of Prompting Techniques; Journal of management Information System; Spring 2001; Vol. 17 N. 4</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Setting up an Internet sales company</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Systematic interview
➢ Semantic interview</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>They analysed the behaviour of job selection professionals.
n = 15</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Number of requirements</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Accepted</td></tr>
<tr><td>9</td><td>Results</td><td>
<table>
<tr><th>Technique</th><th>Mean number of requirements</th><th>Standard deviation</th></tr>
<tr><td>Systematic Interview</td><td>30.8</td><td>16.25</td></tr>
<tr><td>Semantic Interview</td><td>40.93</td><td>12.06</td></tr>
</table>
</td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 11:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation</td></tr>
<tr><td>2</td><td>Reference</td><td>Agarwal, R.; Tanniru, M.; Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation; Journal of Management Information System, M.E. Sharpe; 1990; Vol. 7 N. 1</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Knowledge elicitation from decision-making experts</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Open interview with novice KEs
➢ Structured interview with novice KEs
➢ Open interview with experienced KEs</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>Company directors
n = 10</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Number of rules
2- Number of criteria</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Variable 1 is considered to be the best response variable for this problem.</td></tr>
<tr><td>9</td><td>Results</td><td><table><tr><th>Technique</th><th>Number of rules</th></tr><tr><td>Open interview with novice KEs</td><td>5.2</td></tr><tr><td>Structured interview with novice KEs</td><td>9.9</td></tr><tr><td>Open interview with experienced KEs</td><td>6.1</td></tr></table></td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>

Empirical study 12:

<table>
<tr><th>Item</th><th>Aspect</th><th>Description</th></tr>
<tr><td>1</td><td>Title</td><td>Enhancing Knowledge Elicitation using the Cognitive Interview</td></tr>
<tr><td>2</td><td>Reference</td><td>Woody, J.; Will, R.; Blanton, J.; Enhancing Knowledge Elicitation using the Cognitive Interview; Expert system with application; 1996; Vol. 10 N. 1</td></tr>
<tr><td>3</td><td>Problem Domain</td><td>Recommendation of books based on a set of topics.</td></tr>
<tr><td>4</td><td>Study Type</td><td>Laboratory quasi-experiment</td></tr>
<tr><td>5</td><td>Techniques</td><td>➢ Cognitive interview
➢ Standard interview</td></tr>
<tr><td>6</td><td>Description of experimental subjects</td><td>Librarians
n = 21</td></tr>
<tr><td>7</td><td>Response variables</td><td>1- Number of events</td></tr>
<tr><td>8</td><td>Evaluation of variables</td><td>Accepted</td></tr>
<tr><td>9</td><td>Results</td><td><table><tr><th>Technique</th><th>Number of events</th><th>Standard deviations</th></tr><tr><td>Cognitive interview</td><td>10</td><td>4.650</td></tr><tr><td>Standard interview</td><td>5</td><td>2.974</td></tr></table></td></tr>
<tr><td>10</td><td>Study rating</td><td>Passed</td></tr>
</table>